Attorney Docket No. 5951.010-US

PATENT

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of: Harris *et al.*

Confirmation No. 4949

Serial No.: 09/615,571

Group Art Unit: 1632

Filed: July 13, 2000

Examiner: S.D. Priebe

For: Polypeptides Having Phospholipase B Activity And Nucleic Acids Encoding Same

DECLARATION UNDER 37 C.F.R. § 1.132

Commissioner for Patents
P.O. Box 1450
Alexandria, Virginia 22313-1450

Sir:

I, Randy Berka, do hereby state and declare that

1. I received a Ph.D. in Microbiology and Immunology from the University of Colorado, Health Science Center, Denver, Colorado, in 1983. I have been employed at Novozymes Biotech, Inc., Davis, California, since 1992 where I am currently a Research Fellow.

2. I have read the Office Action dated January 13, 2004 and the Advisory Action dated July 12, 2004 issued in connection with the above-referenced patent application and understand that claims 100, 102-104, 109-111, 114-115, and 120-124 are rejected under 35 U.S.C. § 112, first paragraph, for lack of written description, and claims 100, 102-104, 109-111, 114-115, 117 and 119-124 are rejected under 35 U.S.C. § 112, first paragraph, for lack of enablement. I respectfully disagree with these conclusions.

3. The Office states that the specification does not adequately describe suitable structural, physical and chemical characteristics of the claimed nucleic acids or the enzyme they encode to distinguish them from nucleic acid sequences which are not claimed because the recited structural relationships are arbitrary since neither the specification nor the prior art discloses any definitive relationship between protein function and % identity or homology at either the nucleotide or amino acid level. I respectfully disagree with the Office's statement.

The recited structural relationships of (1) percent identity of the amino acid sequences encoded by the genes, (2) percent homology of the nucleic acid sequences of the genes, and (3) nucleic acid hybridizations under defined stringent conditions to identify complementary strands of genes encoding the same or similar enzyme or protein function are far from

arbitrary. The structural relationships are based on very conservative, logical and rational scientific deductions that are supported by detailed statistical analyses reported in the scientific literature. The claimed relationships are based on the work of Chothia and Lesk (Chothia and Lesk, 1986, *EMBO J.* 5: 823-826) and Bork (Bork *et al.*, 1994, *Curr. Opin. Struct. Biol.* 4: 393-403; Bork *et al.*, 1998, *J. Mol. Biol.* 283: 707-725), who established a solid relationship between amino acid sequence identity and structural identity of homologous proteins. These investigators found that structural divergence was an exponential function of sequence divergence, expressed in terms of the fraction of residues that differ between sequences. The reliability of structural annotation transferred by homology, therefore, depends on the sequence identity of the homologous proteins (Chothia and Lesk, 1986). Since the collective structural properties of proteins (circumscribed by their primary structure) are responsible for their biological activities, proteins that share a high degree of amino acid sequence identity are known with reasonable certainty to possess the same biochemical/biological activities. This deduction is based on the following observations from the literature.

First, there are literally hundreds of reports in which investigators have used nucleic acids probes from one species to clone genes encoding a homologous protein/enzyme from a heterologous source. This approach has been repeatedly employed by us to clone several genes from diverse fungi including a laccase gene (Berka *et al.*, 1997, *Appl. Environ. Microbiol.* 63: 3151-3157), a phytase gene (Berka *et al.*, 1998, *Appl. Environ. Microbiol.* 64: 4423-4427), a mutanase gene (Fuglsang *et al.*, 2000, *J. Biol. Chem.* 275: 2009-2018), and a 5-aminolevulinate synthase gene (Elrod *et al.*, 2000, *Curr. Genet.* 38: 291-298).

Second, homologues of a newly sequenced gene product can be identified *via* database searches using BLAST, Smith-Waterman and other computer algorithms and structure and function assigned to the gene product (Bork *et al.*, 1998). This is based on the concept that sequence similarity implies structural and functional similarity.

Third, with the exponential growth of sequence databases and protein structure databases over the last 20 years, relationships between sequence similarity and functional similarity have emerged, and particular thresholds of sequence conservation and functional similarity have become increasingly apparent. It is clear that functional classification is conserved over a range of sequence similarity and biological/biochemical function diverges only when sequence similarities are low enough that they have no statistical significance. In the present case, a conservative sequence identity threshold (90%) was chosen at which all homologues in a BLAST search of the public databases gave only functionally identical enzymes. I provide seven examples in which BLASTP was used to query a publicly available sequence databases (Protein Information Resource, PIR-NREF, GeneseqP) with different fungal enzymes (amylase, acid protease, glucoamylase, exocellobiohydrolase, endoglucanase, phytase, and lipase) and asked if homologous proteins with at least 90% identity possessed the same biological/biochemical activity. The results summarized in Appendix 1 show clearly and convincingly that proteins with 90% sequence identity are annotated to have the same activity. These deductions are further supported by Wilson and colleagues (Wilson *et al.*, 2000, *J. Mol. Biol.* 297: 233-249) who established a clear relationship between sequence similarity and functional similarity. Wilson *et al.* found that functional identity is conserved down to approximately 40% amino acid sequence identity, and that among proteins that share 50-100% sequence identity, function is conserved in almost all. This observation extends the previous observations of Clothia and Lesk (1986) which compared 32 pairs of homologous proteins and

found that with pairs whose sequence identity is greater than 50%, at least 90% of the residues lie in structurally common cores. It is noteworthy that Wilson *et al.* also found that percent identity is more effective at quantifying functional conservation than probabilistic scores (*P*-values, *E*-scores). Additionally, Clothia and Lesk (1986) note that a protein provides a close structural model for other homologous proteins with which its sequence is at least 50% identical.

The essential feature of the claimed invention is isolated nucleic acids that hybridize to nucleotides 568 to 2045 of SEQ ID NO: 1 under the specified stringency conditions and have a specified percent homology and encode polypeptides with a specific function, *i.e.*, phospholipase B activity. It is well known that hybridization techniques using a known DNA as a probe under specified stringency conditions were conventional in the art at the time of filing. The claim is drawn to nucleic acids all of which must hybridize with nucleotides 568 to 2045 of SEQ ID NO: 1 and must encode a protein with phospholipase B activity.

The logic of the patent examiner's argument that claimed synthetic sequences of 80-90% sequence identity are unusable as probes is flawed, since the claimed proteins derived from the resulting genes must have phoshpholipase B activity. The fact that two sequences that share 80% identity to SEQ ID NO: 1 may have as little as 40% sequence identity to each other is irrelevant as long as the gene they detect shares 80-90% identity with SEQ ID NO: 1 and the corresponding gene product exhibits phospholipase B activity.

A skilled person in the art would not expect substantial variation among species encompassed within the scope of the claims, because the specified hybridization and percent identity conditions set forth in the claims yield structurally similar DNAs and proteins. There are hundreds of papers where genes coding for proteins of similar function were cloned on the basis of hybridization to heterologous probes under a variety of stringency conditions (*e.g.*, Berka *et al.*, 1997, *Appl. Environ. Microbiol.* 63: 3151-3157; Berka *et al.*, 1998, *Appl. Environ. Microbiol.* 64: 4423-4427; Fuglsang *et al.*, 2000, *J. Biol. Chem.* 275: 2009-2018; Elrod *et al.*, 2000, *Curr. Genet.* 38: 291-298; Kraus *et al.*, 1989, *Proc Nat. Acad. Sci. USA* 86: 9193-9197; Sweitzer *et al.*, 1995, *J. Biol. Chem.* 270: 16510-16513; Kraus and Aaronson, 1991, *Methods Enzymol.* 200: 546-556; Kraus and Aaronson, U.S. Patent No. 6,639,060). In such cases, the structurally similar proteins have a high degree of sequence identity. Such cross-hybridization is to a significant extent predictive of gene relatedness, and gene relatedness is in turn predictive of functional similarity. Furthermore, as noted above, from the publications of Chothia and Lesk (1986), Bork (1994 and 1998), and Wilson *et al.* (2000) it is virtually certain that proteins with a high degree of amino acid sequence identity (>50%) have the same biological/biochemical function.

Each of the claimed structural features (percent identity, percent homology, and hybridization) specifies, therefore, a genus of structurally- and functionally-related enzymes having phospholipase B activity.

4. The Office also states that while one of skill in the art can readily envision numerable species of nucleic acid sequences that are at least 90% identical to the recited reference nucleotide sequence and that encode a polypeptide that is at least 90% identical to the recited reference amino acid sequence, one cannot envision which of these also encode a polypeptide with phospholipase B activity.

As I have described above in paragraph 3, the work of Wilson and colleagues (Wilson *et al.*, 2000) demonstrates convincingly that proteins which share 50-100% sequence identity consistently and reliably represent molecules with the same biochemical/biological function.

5. The Office also states that the specification does not provide any information on what amino acid residues are necessary and sufficient for phospholipase B activity or what amino acid sequence modifications, *e.g.*, insertions, deletions and substitutions, would be permissible in a phospholipase B polypeptide that would improve or at least would not interfere with the biological activity or structural features necessary for the biological activity and stability of the protein. Moreover, the Office argues that since there are no other examples of a phospholipase B known that have structural homology with SEQ ID NO: 2, it is not possible to even guess at the amino acid residues which are critical to its structure or function based on sequence conservation. I disagree with this assertion.

John Maynard Smith proposed more than 30 years ago (Smith, 1970, *Nature* 225: 563-564) that the occurrence of functional mutant proteins that differ from wild-type is frequent for evolution to be possible. Since then, numerous evolutionary and mutagenesis studies have supported the assertion that proteins are highly plastic in tolerating amino acid changes (Creighton, 1993, *Proteins* (Freeman, New York); Bowie *et al.*, 1990, *Science* 247: 1306-1310).

Guo *et al.*, 2004, *Proc. Nat. Acad Sci USA* 101: 9205-9210, observed that various residues of a protein are differentially sensitive to substitutions, and that tolerance of the entire protein to random change can be characterized by a probabilistic relationship termed the "*x*-factor." The *x*-factor is broadly defined as the probability that a random amino acid replacement will lead to functional inactivation. Moreover, they determined the *x*-factor to be 34% $\pm$ 6%. Contrary to the Office's contention that random (even conservative) changes in a protein in the absence of structural information would adversely affect folding and/or activity, the findings of Guo *et al.* (2004) support the contrary, *i.e.*, that proteins are generally tolerant to random amino acid substitutions, and the probability of destroying protein function is surprisingly small. Furthermore, Clothia and Lesk (1986) note that the structure of the active site domains may be highly conserved among homologous proteins even when overall amino acid sequence identities are low.

Makiewicz *et al.*, 1994, *J. Mol. Biol.* 240: 421-433, examined 12 or 13 different amino acid substitutions at each residue across 90% of the 360 amino acid *E. coli* lac repressor protein. Reanalysis of their data by Guo *et al.* (2004) revealed an x-factor value of 34% which is identical to the value for random inactivation of human 3-methyladenine DNA glycosylase studied by Guo *et al.* Axe *et al.*, 1998, *Biochem.* 37: 7157-7166, found that 95% of randomly introduced single amino acid substitutions did not lead to inactivated ribonuclease enzyme. Rennell *et al.*, 1991, *J. Mol. Biol.* 222: 67-88, found that approximately 84% of amino acid substitutions in T4 lysozyme did not cause inactivation.

The phospholipase B enzyme described in the present application is novel and represents the first member of a new "genus". However, the enzyme harbors sequence and structural motifs that are well known for enzymes of the phosphoesterase class (which contains not only phospholipase B, but also some phospholipase C, and other phosphomonoesterase enzymes). Using a standard software tool (HMMPFAM) that is well known in the art

(Sonnhammer *et al.*, 1998, *Nucleic Acids Research* 26: 320-322), a profile of conserved amino acid motifs can be generated representing highly conserved sequences in this family. As noted by Guo *et al.* (2004), such highly conserved segments may be critical for enzyme activity or biological function, and they are expected to be less tolerant for substitutions (see Appendix 2 for an example of this analysis).

A skilled person with such information could easily prepare a variant of SEQ ID NO: 2 containing a deletion, insertion, and/or substitution of one or more amino acid residues. The specification on page 7, lines 8-30, provides that conservative amino acid substitutions can be made that do not significantly affect the folding and/or activity of SEQ ID NO: 2 and provide examples of conservative substitutions within the group of basic amino acids, acidic amino acids, polar amino acids, hydrophobic amino acids, aromatic amino acids, and small amino acids. The specification on page 13, line 9, to page 14, line 4, also describes methods for identifying amino acid residues essential to the activity of a phospholipase B.

The Office is, therefore, incorrect in stating that it is not possible to even guess at the amino acid residues which are critical to its structure or function based on sequence conservation.

6. The Office states that Example 2 "does not demonstrate using a nucleic acid of SEQ ID NO: 1 to isolate a claimed nucleic acid from a different source, nor does the specification identify a source, from which one would be able to isolate a claimed nucleic acid, other than *A. oryzae*. More importantly, the claims are not limited to sequences obtainable from a natural source, and the example does not teach how to make a nucleic acid readable on the claims that cannot be found in nature and encodes a different amino acid sequence than SEQ ID NO: 2."

Preliminarily, the specification contains an extensive disclosure of techniques which are well known in the art and indeed routine for persons of ordinary skill for identifying other nucleotides of the present invention. The specification describes methods for preparing and probing DNA libraries (Example 1-2); for isolating nucleic acids encoding the phospholipases (Example 3); for determining cross-hybridization of the nucleic acids encoding phospholipases using (i) nucleotides 568 to 2045 of SEQ ID NO:1, (ii) the cDNA sequence contained in nucleotides 568 to 2045 of SEQ ID NO:1, or (iii) a complementary strand of (i) or (ii) (page 5, line 1, to page 7, line 7); for comparing the percent identity of the deduced amino acid sequences of the phospholipases to amino acids 20 to 464 of SEQ ID NO: 2 using the Clustal method according to Higgins, 1989, *CABIOS* 5: 151-153 (Example 4); for determining the degree of homology between two nucleic acid sequences using the Wilbur-Lipman method according to Wilbur and Lipman, 1983, *Proceedings of the National Academy of Science USA* 80: 726-730 (page 12, line 29, to page 13, line 8); for producing the phospholipases (Example 5); and for purifying the phospholipases and characterizing the properties of the encoded phospholipases (Examples 6-9). A skilled person could easily isolate and identify the claimed nucleic acid sequences using Applicants' disclosure.

The Office indicates that the specification does not demonstrate the isolation of a claimed nucleic acid from a different source using SEQ ID NO: 1 or SEQ ID NO: 2, nor does the specification identify a source, from which one would be able to isolate a claimed nucleic acid, other than *A. oryzae*. I conducted a BLASTP search of several publicly available protein databases (NR, PIRNREF, GENESEQP, SWALL) using SEQ ID NO: 2 as the query sequence to

determine whether SEQ ID NO: 2 could be used to identify homologues that encode a phospholipase subsequent to the filing date of the instant application. The results of the search revealed four proteins having phospholipase activity from *Aspergillus niger*, *Oryza sativa*, and *Burkholderia pseudomallei*:

(1) Accession no. ADF82794: *Aspergillus niger* phospholipase PLP03, Expectancy = 1e-139, Identities = 429/444 (54%).
(2) Accession no. NR 52076602: phospholipase -like protein [*Oryza sativa* (japonica cultivar-group)], Expectancy = 2e-44, Identities = 143/414 (34%).
(3) Accession no. NR 50919526: putative phospholipase [Oryza sativa], Expectancy = 4e-40, Identities = 136/417 (32%).
(4) Accession no. NR 53719489: putative phospholipase [*Burkholderia pseudomallei* K96243], Expectancy = 1e-36, Identities = 130/418 (31%).

These results clearly demonstrate the ability to use SEQ ID NO: 2 to identify proteins having phospholipase activity from other sources. A skilled person would be able to isolate the gene encoding these phospholipases using Applicants' specification.

With regard to man-made variants, a skilled person could easily prepare a variant of SEQ ID NO: 2 containing a deletion, insertion, and/or substitution of one or more amino acid residues. The specification on page 7, lines 8-30, provides that conservative amino acid substitutions can be made that do not significantly affect the folding and/or activity of SEQ ID NO: 2 and provide examples of conservative substitutions within the group of basic amino acids, acidic amino acids, polar amino acids, hydrophobic amino acids, aromatic amino acids, and small amino acids. Amino acid substitutions that do not generally alter the specific activity are described, for example, by H. Neurath and R.L. Hill, 1979, *In*, *The Proteins*, Academic Press, New York. The specification on page 13, line 9, to page 14, line 4, further describes methods for identifying amino acid residues essential to the activity of a phospholipase B. In fact, the findings of Guo *et al.* (2004) support the contention that proteins are generally tolerant to amino acid substitutions, and the probability of destroying protein function is surprisingly small.

7. The Office argues that "it is <u>not</u> routine in the art to screen for multiple substitutions or multiple modifications, as encompassed by the instant claims ..." I disagree with this statement. As of October 1999, a skilled person was able to routinely produce thousands of mutants of SEQ ID NO: 1 through mutagenesis and other techniques and screen the mutants in a short period of time without undue experimentation. See, for example, Christians *et al.*, 1999, *Nature Biotechnology* 17: 259-264; Zocher *et al.*, 1999, *Analytica Chimica Acta* 391: 345-351; Rieger *et al.*, 1999, *Yeast* 15: 973-986; Genome Analysis, A Laboratory Manual, Volume 3, Cloning Systems, Robotic Replication pp. 20-22, Cold Spring Harbor Laboratory Press, 1997; Kell, 1999, *Trends in Biotechnology* 17: 89-91; and Dove, 1999, *Nature Biotechnology* 17: 859-863; Armstrong *et al.*, 1998, *Journal of Biomolecular Screening* 3: 271-275; Eickhoff *et al.*, 1999, *BioMethods* 10: 17-30; and Stevens *et al.*, 1998, *Journal of Biomolecular Screening* 3: 305-311. In addition, the specification provides on page 13, line 9, to page 14, line 4, how to identify essential amino acids in the sequence of SEQ ID NO: 2. A skilled person can, therefore, predict with reasonable statistical accuracy which modifications, if any, would result in a loss of the desired activity/utility.

8. The undersigned declarant declares further that all statements made herein of her own knowledge are true and that all statements made on information and belief are believed to be true and further that these statements are made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize any patent issuing thereon.

Signed this 19th day of November 2004

Randy M. Berka

APPENDIX 1

Demonstration that proteins which share 50-100% amino acid sequence identity are annotated to have the same biochemical/biological function.

A. Query sequence = *Aspergillus niger* glucoamylase (glucan 1,4-alpha-glucosidase (EC 3.2.1.3)). The following BLASTP hits with at least 50% sequence identity are all annotated as glucoamylase enzymes (except those noted as hypothetical or unnamed products):

Abstract	E-value	% Identity
pirnref|NF00073574| Glucoamylase I precursor (EC 3.2.1.3) (Gluca...	0.0	100
pirnref|NF00889958| glucan 1,4-alpha-glucosidase (EC 3.2.1.3) pr...	0.0	100
pirnref|NF00626751| glucan 1,4-alpha-glucosidase (EC 3.2.1.3) pr...	0.0	98
pirnref|NF00626853| Glucoamylase precursor (EC 3.2.1.3) (Glucan ...	0.0	98
pirnref|NF00889947| Glucoamylase precursor (EC 3.2.1.3) [Aspergi...	0.0	97
pirnref|NF01651009| glucoamylase [Aspergillus awamori]	0.0	96
pirnref|NF00626460| Glucoamylase precursor (EC 3.2.1.3) (Glucan ...	0.0	94
pirnref|NF00889975| Glucoamylase precursor (EC 3.2.1.3) (Glucan ...	0.0	94
pirnref|NF01328097| Glucoamylase [Aspergillus niger]	0.0	93
pirnref|NF00626366| preproglucoamylase G2 [Aspergillus niger]	0.0	93
pirnref|NF00889945| glucan 1,4-alpha-glucosidase (EC 3.2.1.3) G2...	0.0	93
pirnref|NF00889968| Glucoamylase-471 [Aspergillus awamori]	0.0	98
pirnref|NF00889964| Glucoamylase-471 [Aspergillus awamori]	0.0	98
pirnref|NF00889951| Glucoamylase-471 (1,4-Alpha-D-Glucan Glucohy...	0.0	98
pirnref|NF00626575| Glucoamylase precursor (EC 3.2.1.3) (Glucan ...	0.0	66
pirnref|NF00494189| Glucoamylase precursor (EC 3.2.1.3) [Talarom...	0.0	61
pirnref|NF00649388| glucan 1,4-alpha-glucosidase (EC 3.2.1.3) pr...	0.0	55
pirnref|NF00647663| Glucoamylase precursor (EC 3.2.1.3) (Glucan ...	0.0	55
pirnref|NF00648280| glucan 1,4-alpha-glucosidase [Neurospora cra...	0.0	55
pirnref|NF01576653| hypothetical protein MG01096.4 [Magnaporthe ...	0.0	53
pirnref|NF01709909| hypothetical protein FG06278.1 [Gibberella z...	e-173	50

B. Query sequence = *Aspergillus niger* aspergillopepsin (acid proteinase/aspartyl protease/preproproctase). The following BLASTP hits with at least 50% sequence identity are all annotated as acid protease/aspartyl protease enzymes (except those noted as hypothetical or unnamed products):

Abstract	E-value	% Identity

pirnref|NF00626537| Aspergillopepsin A precursor (EC 3.4.23.18) ...	0.0	100
pirnref|NF00626722| aspergillopepsin I (EC 3.4.23.18) precursor ...	0.0	99
pirnref|NF00918479| Aspergillopepsin A precursor (EC 3.4.23.18) ...	0.0	99
pirnref|NF00626425| Preproproctase B precursor [Aspergillus niger]	0.0	96
pirnref|NF00889972| Aspergillopepsin A precursor (EC 3.4.23.18) ...	0.0	96
pirnref|NF00626729| Aspergillopepsin [Aspergillus phoenicis]	0.0	99
pirnref|NF00627288| Aspergillopepsin i (EC 3.4.23.18) [Aspergill...	e-163	71
pirnref|NF00626684| Aspergillopepsin A precursor (EC 3.4.23.18) ...	e-155	67
pirnref|NF00626584| aspergillopepsin O [Aspergillus oryzae]	e-155	67
pirnref|NF00626993| Propenicillopepsin-JT2 precursor [Penicilliu...	e-152	67
pirnref|NF00176292| Putative aspartic protease [Emericella nidul...	e-145	65
pirnref|NF00889953| aspergillopepsin I (EC 3.4.23.18) [Aspergill...	e-144	81
pirnref|NF00627293| Aspergillopepsin F precursor (EC 3.4.23.18) ...	e-142	66
pirnref|NF01463777| acid proteinase [Monascus purpureus]	e-141	63
pirnref|NF00627188| Aspartic proteinase [Penicillium roquefortii]	e-138	63
pirnref|NF00626580| Aspartic proteinase II-1 [Aspergillus oryzae]	e-137	65
pirnref|NF01229506| Aspartic Proteinase [Aspergillus oryzae]	e-129	70
pirnref|NF00627002| Prepropenicillopepsin-JT3 precursor [Penicil...	e-129	58
pirnref|NF00626995| Penicillopepsin (EC 3.4.23.20) (Peptidase A)...	e-127	68
pirnref|NF00626992| penicillopepsin (EC 3.4.23.20) [Penicillium ...	e-127	68
pirnref|NF00747468| Aspartic proteinase precursor (EC 3.4.23.-) ...	e-104	49
pirnref|NF00646517| Endothiapepsin precursor (EC 3.4.23.22) (Asp...	e-103	50
pirnref|NF01576663| hypothetical protein MG02898.4 [Magnaporthe ...	e-103	49
pirnref|NF00646493| Endothiapepsin [Cryphonectria parasitica]	e-102	55

C. Query sequence = *Aspergillus oryzae* α-amylase (AMY1, Taka-amylase, *amyA*). The following BLASTP hits with at least 50% sequence identity are all annotated as α-amylase enzymes (except those noted as hypothetical or unnamed products):

Abstract	E-value	% Identity
pirnref|NF00626669| Alpha-amylase A precursor (EC 3.2.1.1) (Taka...	0.0	100
pirnref|NF01651008| alpha-amylase [Aspergillus awamori]	0.0	99
pirnref|NF00626583| unnamed protein product [Aspergillus oryzae]	0.0	100
pirnref|NF01544944| alpha-amylase [Aspergillus kawachii]	0.0	100
pirnref|NF00626750| alpha-amylase (EC 3.2.1.1) precursor [Asperg...	0.0	99
pirnref|NF00626854| Alpha-amylase precursor (EC 3.2.1.1) (1,4-al...	0.0	99
pirnref|NF00626612| Taka-amylase A (EC 3.2.1.1) (Alpha-amylase) ...	0.0	99

pirnref|NF00625791| Taka-amylase A (EC 3.2.1.1) (Alpha-amylase) ...	0.0	99
pirnref|NF00626368| alpha-amylase-precursor [Aspergillus niger]	0.0	99
pirnref|NF00889948| Alpha-amylase B precursor (EC 3.2.1.1) (1,4-...	0.0	99
pirnref|NF00626646| alpha-amylase (EC 3.2.1.1) precursor [Asperg...	0.0	99
pirnref|NF00626648| Taka-amylase A (Taa-G1) precursor [Aspergill...	0.0	99
pirnref|NF00626351| alpha-amylase-precursor [Aspergillus niger]	0.0	99
pirnref|NF00889969| Alpha-amylase A precursor (EC 3.2.1.1) (1,4-...	0.0	99
pirnref|NF00626590| alpha-amylase (EC 3.2.1.1) precursor [Asperg...	0.0	99
pirnref|NF00626638| Taka Amylase [Aspergillus oryzae]	0.0	100
pirnref|NF00626642| alpha-amylase (EC 3.2.1.1) [Aspergillus oryzae]	0.0	97
pirnref|NF00176034| Alpha-amylase AmyA [Emericella nidulans]	0.0	69
pirnref|NF00073571| Acid-stable alpha-amylase [Aspergillus kawac...	0.0	67
pirnref|NF01651007| alpha-amylase [Aspergillus awamori]	0.0	68
pirnref|NF00626487| alpha-amylase (EC 3.2.1.1) [Aspergillus niger]	0.0	67
pirnref|NF00626518| Acid alpha-amylase (EC 3.2.1.1) (1,4-alpha-D...	0.0	66
pirnref|NF00176203| Alpha-amylase [Emericella nidulans]	0.0	63
pirnref|NF01752634| alpha-amylase precursor [Lipomyces starkeyi]	0.0	60
pirnref|NF00756572| unnamed protein product [Thermomyces lanugin...	0.0	60
pirnref|NF00409123| Lipomyces kononenkoae subsp. spencermartinsi...	e-180	57
pirnref|NF00186159| Alpha-amylase 1 precursor (EC 3.2.1.1) (1,4-...	e-180	56
pirnref|NF00490302| Alpha-amylase 2 precursor (EC 3.2.1.1) (1,4-...	e-167	56
pirnref|NF00490307| Alpha-amylase 1 precursor (EC 3.2.1.1) (1,4-...	e-159	56
pirnref|NF00490293| alpha-amylase (EC 3.2.1.1) precursor [Debary...	e-158	55
pirnref|NF00490296| alpha-amylase [Debaryomyces occidentalis]	e-158	54
pirnref|NF00155569| alpha-amylase [synthetic construct]	e-153	53

D. Query sequence = *Hypocrea jecorina (Trichoderma reesei)* exocellobiohydrolase I (CBH1, exoglucanase, cellobiohydrolases, 1,4,-β-glucan cellobiohydrolase). The following BLASTP hits with at least 50% sequence identity are all annotated as exocellobiohydrolase enzymes (except those noted as hypothetical or unnamed products):

Abstract	E-value	% Identity
pirnref|NF00769949| Exoglucanase I precursor (EC 3.2.1.91) (Exoc...	0.0	100
pirnref|NF01042178| cellulose 1,4-beta-cellobiosidase (EC 3.2.1....	0.0	100
pirnref|NF00494383| Exoglucanase I precursor (EC 3.2.1.91) (Exoc...	0.0	100
pirnref|NF01470257| cellobiohydrolase I [Trichoderma viride]	0.0	99

pirnref|NF00756631| Cellobiohydrolase I [Trichoderma viride]	0.0	95
pirnref|NF00756635| Exoglucanase I precursor (EC 3.2.1.91) (Exoc...	0.0	94
pirnref|NF00494360| Cellobiohydrolase I [Hypocrea jecorina]	0.0	100
pirnref|NF00494368| 1,4-Beta-D-Glucan Cellobiohydrolase I [Hypoc...	0.0	99
pirnref|NF00494367| 1,4-Beta-D-Glucan Cellobiohydrolase I [Hypoc...	0.0	99
pirnref|NF00494366| 1,4-Beta-D-Glucan Cellobiohydrolase I [Hypoc...	0.0	99
pirnref|NF01524434| Exocellobiohydrolase I [Hypocrea jecorina]	0.0	99
pirnref|NF00494322| 1,4-Beta-D-Glucan Cellobiohydrolase Cel7 [Hy...	0.0	98
pirnref|NF01524433| Exocellobiohydrolase I [Hypocrea jecorina]	0.0	97
pirnref|NF00756590| Cellobiohydrolase [Hypocrea lixii]	0.0	80
pirnref|NF01265187| unnamed protein product [Acremonium thermoph...	0.0	63
pirnref|NF01265191| unnamed protein product [Chaetomidium pingtu...	0.0	62
pirnref|NF00835152| Xylanase/cellobiohydrolase precursor (EC 3.2...	0.0	62
pirnref|NF01288915| unnamed protein product [Exidia glandulosa]	0.0	61
pirnref|NF00626501| 1,4-beta-D-glucan cellobiohydrolase B precur...	0.0	59
pirnref|NF01266275| unnamed protein product [Chaetomium thermoph...	0.0	58
pirnref|NF01489984| Hypothetical protein [Neurospora crassa]	0.0	59
pirnref|NF01258404| unnamed protein product [Scytalidium thermop...	0.0	57
pirnref|NF00992461| 1,4-beta-D-glucan-cellobiohydrolyase (EC 3.2...	0.0	58
pirnref|NF00756327| Cellulase (EC 3.2.1.91) [Humicola grisea]	0.0	57
pirnref|NF01286453| unnamed protein product [Thermoascus auranti...	0.0	65
pirnref|NF00756321| Exoglucanase I precursor (EC 3.2.1.91) (Exoc...	0.0	57
pirnref|NF00801194| Cellulase CEL7A [Lentinula edodes]	0.0	60
pirnref|NF01257476| unnamed protein product [Thielavia australie...	0.0	56
pirnref|NF00625663| Exoglucanase I precursor (EC 3.2.1.91) (Exoc...	0.0	58
pirnref|NF00962307| Cellobiohydrolase I [Thermoascus aurantiacus]	e-180	64
pirnref|NF00856066| Cellobiohydrolase [Thermoascus aurantiacus]	e-179	64
pirnref|NF00959024| Cellobiohydrolase I catalytic domain (EC 3.2...	e-179	64
pirnref|NF01709668| GUXC_FUSOX Putative exoglucanase type C prec...	e-179	57
pirnref|NF01053514| Cellobiohydrolase C [Aspergillus oryzae]	e-179	63
pirnref|NF00755639| Putative exoglucanase type C precursor (EC 3...	e-179	57

pirnref|NF00626413| 1,4-beta-D-glucan cellobiohydrolase A precur...	e-177	65
pirnref|NF00627000| Exoglucanase I precursor (EC 3.2.1.91) (Exoc...	e-177	57
pirnref|NF01696049| cellobiohydrolase C [Gibberella zeae]	e-176	57
pirnref|NF01288914| unnamed protein product [Exidia glandulosa]	e-176	63
pirnref|NF00646477| Exoglucanase I precursor (EC 3.2.1.91) (Exoc...	e-174	62
pirnref|NF01581876| hypothetical protein MG06834.4 [Magnaporthe ...	e-174	62
pirnref|NF00648798| Exoglucanase 1 precursor (EC 3.2.1.91) (Exoc...	e-170	56
pirnref|NF01265188| unnamed protein product [Trichophaea saccata]	e-169	60
pirnref|NF00992462| 1,4-beta-D-glucan-cellobiohydrolyase (EC 3.2...	e-168	60
pirnref|NF01053511| Cellobiohydrolase D [Aspergillus oryzae]	e-163	60
pirnref|NF00731508| cellulose 1,4-beta-cellobiosidase (EC 3.2.1....	e-163	54
pirnref|NF00731784| Cellulase precursor [Irpex lacteus]	e-162	52
pirnref|NF00733334| Exoglucanase precursor (EC 3.2.1.91) (Exocel...	e-162	54
pirnref|NF00731509| cellulose 1,4-beta-cellobiosidase (EC 3.2.1....	e-161	54
pirnref|NF00731785| Exocellulase precursor [Irpex lacteus]	e-160	53

E. Query sequence = *Hypocrea jecorina (Trichoderma reesei)* endoglucanase I I (EG1, endo-1,4-β-glucanase, 1,4,-β-glucan glucanhydrolase). The following BLASTP hits with at least 50% sequence identity are all annotated as endoglucanase enzymes (except those noted as hypothetical or unnamed products):

Abstract	E-value	% Identity
pirnref|NF00494331| Endoglucanase EG-1 precursor (EC 3.2.1.4) (E...	0.0	100
pirnref|NF01407727| Endoglucanase I [Trichoderma viride]	0.0	99
pirnref|NF00756647| Endoglucanase EG-1 precursor (EC 3.2.1.4) (E...	0.0	94
pirnref|NF00756639| Endoglucanase I [Trichoderma viride]	0.0	93
pirnref|NF00154649| ENDO II [synthetic construct]	0.0	97
pirnref|NF00494347| Endoglucanase I [Hypocrea jecorina]	0.0	100
pirnref|NF00793302| unnamed protein product [Talaromyces emersonii]	e-121	55
pirnref|NF00626671| Endo-1,4-beta-glucanase (EC 3.2.1.4) [Asperg...	e-107	51

F. Query sequence = *Coprinus cinereus* laccase (polyphenoloxidase, bilirubin oxidase, multicopper oxidase). The following BLASTP hits with at least 50% sequence identity are all annotated as laccase enzymes (except those noted as hypothetical or unnamed products):

Abstract	E-value	% Identity

pirnref|NF00733435| Laccase 2 (EC 1.10.3.2) [Coprinopsis cinerea]	0.0	100
pirnref|NF00733482| Laccase 3 (EC 1.10.3.2) [Coprinopsis cinerea]	0.0	78
pirnref|NF01638355| laccase 3 [Coprinopsis cinerea]	0.0	78
pirnref|NF01386173| Laccase 4 (EC 1.10.3.2) [Pleurotus sajor-caju]	0.0	64
pirnref|NF00731916| Laccase 2 precursor (EC 1.10.3.2) (Benzenedi...	0.0	64
pirnref|NF01386176| Laccase 5 (EC 1.10.3.2) [Pleurotus sajor-caju]	0.0	66
pirnref|NF00731901| Bilirubin oxidase (Laccase) [Pleurotus ostre...	0.0	64
pirnref|NF01567552| laccase [Pleurotus ostreatus]	0.0	64
pirnref|NF01386172| Laccase 2 (EC 1.10.3.2) [Pleurotus sajor-caju]	0.0	67
pirnref|NF01461741| laccase [Rigidoporus microporus]	0.0	65
pirnref|NF01461740| laccase [Rigidoporus microporus]	0.0	65
pirnref|NF01386175| Laccase 1 (EC 1.10.3.2) [Pleurotus sajor-caju]	0.0	62
pirnref|NF00731925| Laccase 1 precursor (EC 1.10.3.2) (Benzenedi...	0.0	62
pirnref|NF01637249| laccase [Pleurotus ostreatus] [Pleurotus pul...	0.0	62
pirnref|NF00427931| Laccase precursor [Funalia trogii]	0.0	63
pirnref|NF00758114| Laccase precursor (EC 1.10.3.2) [basidiomyce...	0.0	63
pirnref|NF00939119| Polyphenoloxidase (EC 1.10.3.2) (Laccase 1) ...	0.0	63
pirnref|NF00232919| Polyphenoloxidase (EC 1.10.3.2) (Laccase 1) ...	0.0	63
pirnref|NF01689789| laccase [Pleurotus ostreatus]	0.0	62
pirnref|NF00993008| Laccase 2 (EC 1.10.3.2) [Trametes pubescens]	0.0	64
pirnref|NF00732955| Laccase (EC 1.10.3.2) [Schizophyllum commune]	0.0	63
pirnref|NF00731988| Laccase precursor (EC 1.10.3.2) (Benzenediol...	0.0	63
pirnref|NF00731957| laccase (EC 1.10.3.2) A [Trametes versicolor]	0.0	63
pirnref|NF00731968| ligninolytic phenoloxidase (EC 1.10.-.-) 2 p...	0.0	63
pirnref|NF01057965| Laccase 2 [Trametes versicolor]	0.0	64
pirnref|NF00044343| Laccase 2 precursor (EC 1.10.3.2) (Benzenedi...	0.0	64
pirnref|NF00731635| Laccase precursor (EC 1.10.3.2) (Benzenediol...	0.0	63
pirnref|NF00059532| Laccase precursor (EC 1.10.3.2) [Coriolus ve...	0.0	63
pirnref|NF00731977| laccase I [Trametes versicolor]	0.0	64
pirnref|NF00788162| Laccase [Pycnoporus coccineus]	0.0	63
pirnref|NF00731989| ligninolytic phenoloxidase [Trametes hirsuta]	0.0	63
pirnref|NF00788163| Laccase precursor [Pycnoporus coccineus]	0.0	63
pirnref|NF00945391| unnamed protein product [unidentified]	0.0	64
pirnref|NF00909761| Laccase (Laccase 1) [Lentinula edodes]	0.0	64
pirnref|NF00044346| Laccase 1 precursor (EC 1.10.3.2) (Benzenedi...	0.0	63
pirnref|NF00731959| Laccase 2 precursor (EC 1.10.3.2) (Benzenedi...	0.0	64

pirnref|NF00964375| Laccase III (EC 1.10.3.2) [Trametes versicolor]	0.0	63
pirnref|NF00731973| Laccase precursor (EC 1.10.3.2) [Trametes ve...	0.0	63
pirnref|NF00801188| Laccase B precursor (EC 1.10.3.2) [Trametes ...	0.0	63
pirnref|NF01012946| Laccase [Trametes versicolor]	0.0	64
pirnref|NF00050826| Laccase LCC3-1 (EC 1.10.3.1) [Polyporus cili...	0.0	63
pirnref|NF00427925| Laccase (EC 1.10.3.2) [Coriolopsis gallica]	0.0	63
pirnref|NF01470431| laccase [Trametes sp. I-62]	0.0	63
pirnref|NF01470433| laccase [Trametes sp. I-62]	0.0	63
pirnref|NF00466979| Phenoloxidase (EC 1.10.3.2) [Trametes sp. I-62]	0.0	62
pirnref|NF01470432| laccase [Trametes sp. I-62]	0.0	62
pirnref|NF00466977| Phenoloxidase (EC 1.10.3.2) [Trametes sp. I-62]	0.0	62
pirnref|NF01470430| laccase [Trametes sp. I-62]	0.0	62
pirnref|NF00801189| Laccase 1 (EC 1.10.3.2) [Trametes versicolor]	0.0	63

G. Query sequence = *Thermomyces lanuginosus (Humicola lanuginose)* lipase. The following BLASTP hits with at least 50% sequence identity are all annotated as lipase enzymes (except those noted as hypothetical or unnamed products):

Abstract	E-value	% Identity
pirnref|NF00756570| Lipase precursor (EC 3.1.1.3) (Triacylglycer...	e-171	100
pirnref|NF00756566| LIPASE [Thermomyces lanuginosus]	e-159	100
pirnref|NF00756565| Lipase (E.C. 3.1.1.3) (Triacylglycerol Acylh...	e-159	100
pirnref|NF01188178| Lipase [Thermomyces lanuginosus]	e-158	99
pirnref|NF01102488| unnamed protein product [Talaromyces thermop...	e-153	88
pirnref|NF01111633| unnamed protein product [Thermomyces ibadane...	e-136	78
pirnref|NF01114192| unnamed protein product [Talaromyces emersonii]	5e-97	61
pirnref|NF01105706| unnamed protein product [Talaromyces byssoch...	2e-89	57
pirnref|NF00626823| unnamed protein product [Aspergillus tubinge...	8e-77	50
pirnref|NF00158307| unnamed protein product [unidentified]	1e-76	50

APPENDIX 2. HMMPFAM Analysis of Aspergillus oryzae phospholipase

```
Raw Viewer
      hercules.ngcsn.netRaw ViewerPortal
      New search
      Help


hmmpfam - search a single seq against HMM database
HMMER 2.1.1 (Dec 1998)
Copyright (C) 1992-1998 Washington University School of Medicine
HMMER is freely distributed under the GNU General Public License (GPL).
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
HMM file:                 /usr/novo/databases/online/bf_biocofe/pfam/Pfam.pfl
Sequence file:
/usr/novo/projects/biocofe/tmp/20041118_230717_7095.seq
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Query:  unknown,  464 bases, A11C7610 checksum.

Scores for sequence family classification (score includes all domains):
Model           Description                                 Score    E-value  N
--------        -----------                                 -----    ------- ---
Phosphoesterase Phosphoesterase family                      198.0    1.5e-55   1

Parsed for domains:
Model           Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
--------        ------- ----- -----    ----- -----      -----  -------
Phosphoesterase   1/1      51   447 ..     1   526 []   198.0  1.5e-55

Alignments of top-scoring domains:
Phosphoesterase: domain 1 of 1, from 51 to 447: score 198.0, E = 1.5e-55
                   *->ieHvVilmqENRSFDhyfGtls..gvrgeidavse.esnpl.fsDpn
                      +e++V l+ ENRSFD+++G  +++g++++i++  ++   n   + sDp+
    unknown,    51    VENIVWLILENRSFDNILGGVRrqGLDNPINN--GpFCNYKnASDPS 95


                   slkiqfgkpvwesqvvqggwdpdtgasfqalenqPFrfndttegkPllag
                   s k +       +s +       +d+ +s      +  ++  + ++g  +++g
    unknown,    96 SGKYCTQAKDYDSVF------NDPDHSVTGNNLEFYGTYTPNNG-AIASG 138

                   frvqdlnHswydphsawngGr.nDrWlgadaettakavsgpqvMgyfkrs
                    +v+d+          ++ ++  nD+         + a+   + qvMgy++++
    unknown,   139 KVVADQS--------GFLNAQlNDY------PKLAPEEATRQVMGYYTEE 174

                   dipvyLwaLAdeFtlcDnyFcsvpGpTqPNRlyllsGtspfyDdSAKhvS
                   ++p +  +L+deFt ++ +F+ vpGpT+PNRl  l+Gt+
    unknown,   175 EVPTL-VDLVDEFTTFNSWFSCVPGPTNPNRLCALAGTA----------- 212

                   VegmdgdGtlkiandsPasaDGPKFvDGlTPDfYavntmnPpYqpssvps
                       g G+
    unknown,   213 ----AGHGK----------------------------------------- 217

                   knGgpvlanpsiqkpplqgfnwsTipdrLdekGvsWgiYqeklpgtlyqg
                    n++ +l             g + + i++   ekGvsW +Y ++ +g +++
    unknown,   218 -NDDDFLN---------YGISSKSIFEAANEKGVSWLNY-DGTNGEFEPD 256

                   klgfnqyvqyfkqnanplnYwkeysnshaplrladsrklkavrkfhiydl
```

```
               +l+f                                 + +++ +++ +++
unknown,   257 SLFF--------------------------------TYVNQTSRSNVVPV 274

               ssFkkDvkngkLPqVSfiiPRYfDlllnganDeYmHPghdviaaGdkwik
               ++F++D+  g LP+ S+i P      +  +n++  HP+ +v  +G++++k
unknown,   275 ENFFQDAYLGVLPKFSYINP---SCCGTNTNSM--HPTGNV-SYGEVFVK 318

               evleaLlanpqvWnktllivtYDEngGfyDHVppPvapvpnp.glvtvsd
               ++++a+++ pq W ktll++tYDE gGfyDHVppP a +p+   ++
unknown,   319 QIYDAIRQGPQ-WDKTLLFITYDETGGFYDHVPPPLAVRPDNlTYT---- 363

               idav.pGpgpfnifgfyGLGpRVPtlvISPwPskgGtvdhepnGPtpss.
                ++++ G   +    f++LG+R Pt vISP+ sk+G++ ++  G  p ++
unknown,   364 -ETAkNGQKYTL--HFDRLGGRMPTWVISPY-SKKGYIEQY--GTDPVTg 407

               ...tfdHtSvLafiekrFgLpsLpnisawrdavagdltst<-*
               ++  ++ tSvL+++ +++++   +  +a+  ++   +  t
unknown,   408 kpaPYSATSVLKTLGYLWDIEDFTPRVAHSPSFDHLIGTT    447
```

The EMBO Journal vol.5 no.4 pp.823–826, 1986

The relation between the divergence of sequence and structure in proteins

Cyrus Chothia[1] and Arthur M.Lesk[2]

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, and [1]Christopher Ingold Laboratory, University College London, 20 Gordon Street, London WC1H 0AJ, UK

[2]Permanent address: Fairleigh Dickinson University, Teaneck-Hackensack Campus, Teaneck, NJ 07666, USA

Communicated by M.F.Perutz

Homologous proteins have regions which retain the same general fold and regions where the folds differ. For pairs of distantly related proteins (residue identity ~20%), the regions with the same fold may comprise less than half of each molecule. The regions with the same general fold differ in structure by amounts that increase as the amino acid sequences diverge. The root mean square deviation in the positions of the main chain atoms, Δ, is related to the fraction of mutated residues, H, by the expression: $\Delta(\text{Å}) = 0.40\ e^{1.87H}$.

Key words: evolution/protein homology/model building

Introduction

The comparative analysis of the structures of related proteins can reveal the effects of the amino acid sequence changes that have occurred during evolution (Perutz *et al.*, 1965). Previous work on individual protein families has shown that mutations, insertions and deletions produce changes in three-dimensional structure (Almassy and Dickerson, 1978; Lesk and Chothia, 1980, 1982, 1986; Greer, 1981; Chothia and Lesk, 1982, 1984; Read *et al.*, 1984). Here we report a systematic comparison of structures from eight different protein families. This shows that the extent of the structural changes is directly related to the extent of the sequence changes.

In the work reported here we used the atomic coordinates of 25 proteins (Table I). All these structures have been determined at high resolution (1.4–2.0Å) and refined. The errors in their co-ordinates are 0.15–0.20Å (see references given in Table I). The 25 proteins represent eight different protein families and provide 32 pairs of homologous structures.

Methods and Results

The conserved structural cores and the variable regions of homologous proteins

The structures of homologous proteins can be divided into those regions in which the general fold of the polypeptide chains is very similar and those where it is quite different. In comparing protein structures it is useful to separate the parts that have similar folds from those where the folds differ. We did this using the following quantitative procedure: (i) the main-chain atoms of major elements of secondary structure — helices or two adjacent strands of β-sheet — were individually superposed; and (ii) each superposition was then extended to include additional atoms at both ends. The extension was continued as long as the deviations in the positions of the atoms in the last residue included were no greater than 3 Å. This procedure defined the segments that

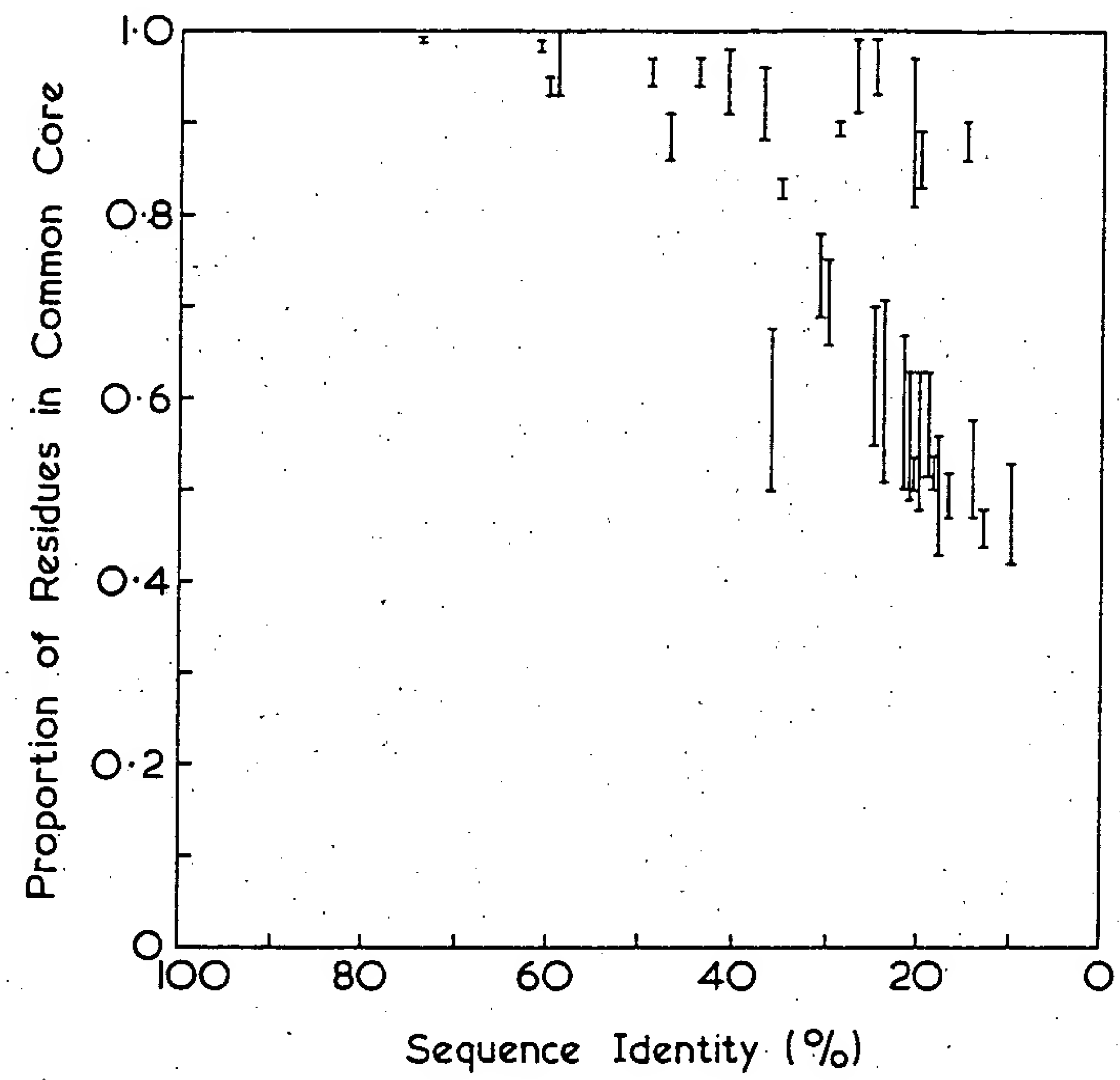


Fig. 1. Size of common cores as a function of protein homology. If two proteins of length n_1 and n_2 have c residues in the common core, the fractions of each sequence in the common core are c/n_1 and c/n_2. We plot these values, connected by a bar, against the residue identity of the core (see Table II).

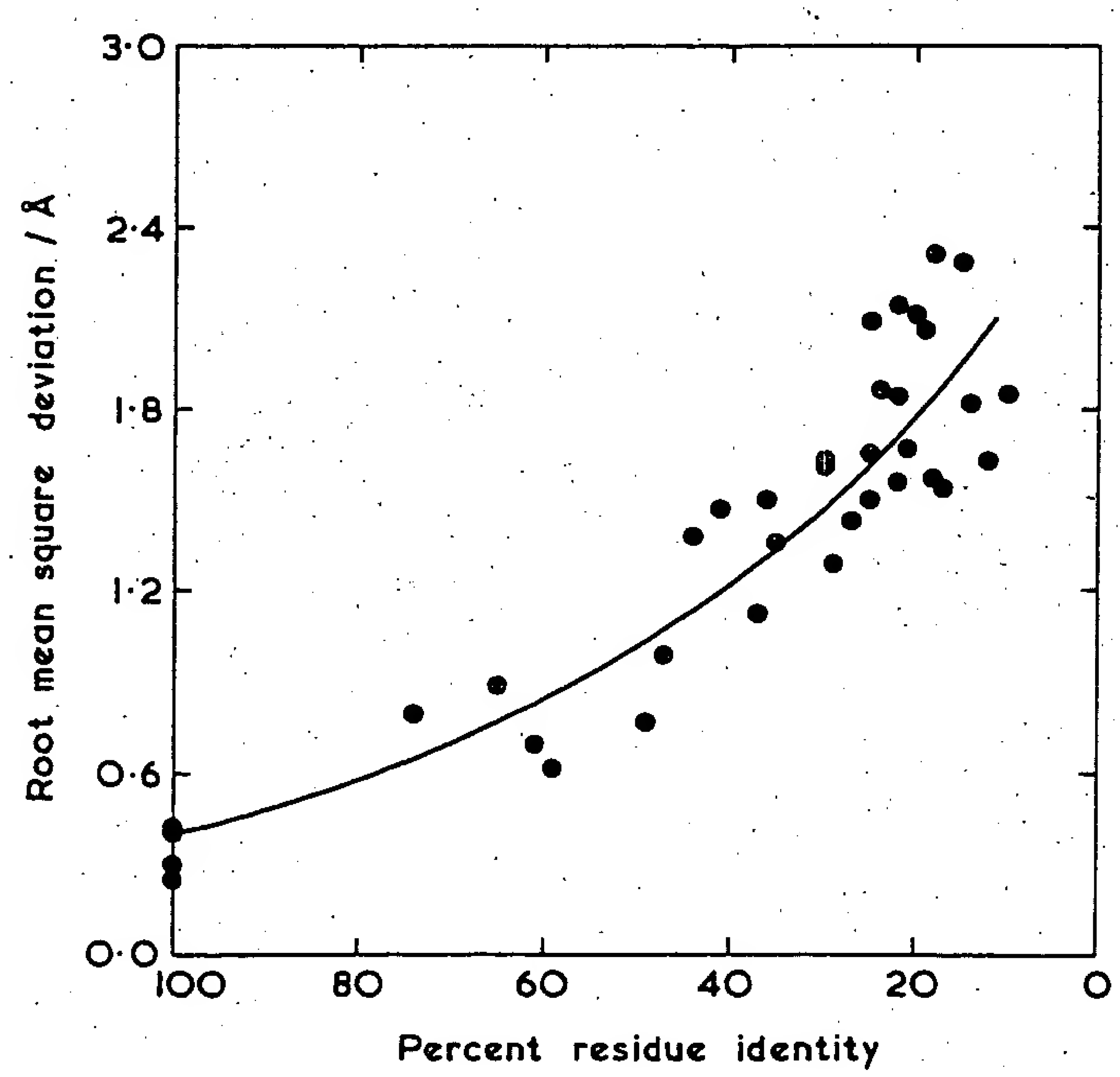


Fig. 2. The relation of residue identity and the r.m.s. deviation of the backbone atoms of the common cores of 32 pairs of homologous proteins (see Table II).

Table I. Homologous proteins determined at high resolution

Family	Protein	Abbreviation	Structure analysis: Resolution (Å)	Structure analysis: R factor, %	Reference
Globins (deoxy)	Human α subunit	HHBα	1.74	16	Fermi *et al.*, 1984
	Human β subunit	HHBβ			
	Sperm whale myoglobin	1MBD	1.40	14	Phillips, 1980
	Erythrocruorin	1ECD	1.40	18	Steigemann and Weber, 1979
Cytochromes	Tuna c	3CYT	1.50	17	Takano and Dickerson, 1982
	Rice embryo c	1CCR	1.50	19	Ochi *et al.*, 1983
	Bacterial c_2	3C2C	1.68	17	Bhatia, 1981
	Bacterial c_{551}	351C	1.60	19	Matsuura *et al.*, 1982
Serine protease	Bovine γ-chymotrypsin	2GCH	1.90	18	Cohen *et al.*, 1981
	Bovine trypsin	3PTP	1.50	16	Chambers and Stroud, 1979
	S. griseus protease A	2SGA	1.80	14	Sielecki *et al.*, 1979
	S. griseus protease B	3SGB	1.80	14	Read *et al.*, 1983
Dihydrofolate reductase	*L. casei*	3DFR	1.70	15	Bolin *et al.*, 1982
	E. coli	4DFR	1.70	17	Bolin *et al.*, 1982
Cu-electron transport proteins	Bacterial azurin	1AZA	2.00	19	Norris *et al.*, 1983
	Poplar leaf plastocyanin	1PCY	1.60	17	Guss and Freeman, 1983
Sulphydryl protease	Papaya papain	PAP	1.65	16	Kamphuis *et al.*, 1985
	Kiwifruit actinidin	2ACT	1.70	17	Baker, 1980
Lysozyme	Human	1LZ1	1.50	18	Artymiuk and Blake, 1981
	Hen egg white	LZHE	1.60		Grace, 1979
Immunoglobulin domains	Vλ (RHE)	2RHE	1.60	15	Furey *et al.*, 1983
	Vλ (KOL)	KLVL	1.90	19	Marquart *et al.*, 1980
	Vγ (KOL)	KLVH			
	Cλ (KOL)	KLCL			
	Cγ_1 (KOL)	KLCH			

Except for hen egg lysozyme and papain, atomic coordinates were obtained from the protein data bank (Bernstein *et al.*, 1977).

have the same fold in both proteins. They include major elements of secondary structure and peptides that form the active site. We call the collection of such regions the 'common core'. The residues outside the common core are in peripheral elements of secondary structure, in the loops between major elements of secondary structure, at the ends of helices, or in strands at the edges of β-sheets (Lesk and Chothia, 1980, 1982; Greer, 1981; Chothia and Lesk, 1982, 1984; Read *et al.*, 1984).

The results of comparing the 32 pairs of homologous proteins are given in Table II. Pairs whose sequence identity is >50% have 90% or more of the residues of the individual structures within the common cores. Pairs whose residue identity drops to about 20% have common cores that contain betwen 42% and 98% of the residues of individual structures (Table II, Figure 1). Proteins built of β-sheets are at the bottom of this range and proteins built of α-helices are at the top. Compared with helical proteins, β-sheet proteins contain proportionally fewer residues within secondary structures and more in loops, the regions particularly susceptible to local refolding when sequences change.

Structural divergence in the common cores of homologous proteins

Although the core regions retain a common fold, they do undergo structural change as their sequences diverge. Mutations at the interfaces between secondary structures produce changes in the geometry of packing and, in the case of β-sheets, limited local changes in backbone conformation (Lesk and Chothia, 1980, 1982, 1986; Chothia and Lesk, 1982, 1984; Read *et al.*, 1984). The overall extent of the structural divergence of two homologous proteins can be measured by optimally superposing the common cores and calculating the root mean square difference in the positions of their main-chain atoms, Δ. For the 32 homologous pairs of proteins in Table II the values of Δ vary between 0.62 and 2.31 Å (Table II).

The exact value of Δ is, of course, dependent upon the procedure used to define the common cores of homologous proteins. Inspection of the regions not in the common cores shows that they usually have very different conformations. This is especially true of the larger loops. Thus modification of the procedure used here to define the common cores would only produce marginal differences.

Essentially similar results are obtained if, in place of a core derived for each individual homologous pair, we use a core common to all members of a family. For example, in the cytochromes c(rice), c(tuna), c_2 and c_{551}, a 48-residue core is common to all four structures (Chothia and Lesk, 1984). Superpositions of this core in the four structures give the Δ values listed in Table III. Compared with the Δ values for individual core comparisons, these Δ values are somewhat smaller for closely related pairs (in these cases the family core is smaller and more homologous than the pair core), but nearly equal for distantly related pairs (Table III).

The contribution to Δ from experimental error and from differences in molecular environment can be estimated from the comparison of proteins whose structures have been accurately determined in different crystal forms, or in crystals that have more than one molecule in the asymmetric unit. The values of Δ for five such proteins are between 0.25 and 0.40 Å (Table II). The mean is 0.33 Å: one half to one seventh of the Δ values reported here for homologous proteins.

Table II. Common cores of homologous proteins: size, fit and residue identity

Family	Protein pair[a]	Residues in protein pair	Residues in core	r.m.s. difference in core (Å)	Percentage of core residues that are the same in both structures
Globin	HHBα:HHBβ	141:146	137	1.38	44
	HHBα:1MBD	141:153	139	1.43	27
	HHBα:1ECD	151:136	122	2.28	15
	HHBβ:1MBD	146:153	143	1.50	25
	HHBβ:1ECD	146:136	121	2.11	20
	1MBD:1ECD	153:136	132	1.67	21
Cytochrome c	3CYT:1CCR	103:111	103	0.62	59
	3CYT:3C2C	103:112	99	1.13	37
	3CYT:351C	103:82	57	1;65	25
	1CCR:3C2C	111:112	101	1.47	41
	1CCR:351C	111:82	58	1.86	24
	3C2C:351C	112:82	56	1.50	36
Serine protease	2GCH:3PTP	236:222	203	0.99	47
	2GCH:2SGA	236:181	114	2.09	25
	2GCH:3SGB	236:185	116	2.14	22
	3PTP:2SGA	221:181	112	1.84	22
	3PTP:3SGB	222:185	116	2.06	19
	2SGA:3SGB	181:185	172	0.89	65
Immunoglobulin domain	2RHE:KLVL	110:110	108	0.80	74
	2RHE:KLVH	110:125	83	1.63	30
	2RHE:KLCL	110:101	55	1.57	18
	2RHE:KLCH	110:99	48	1.47	13
	KLVL:KLVH	110:125	86	1.61	30
	KLVL:KLCL	110:101	55	1.56	22
	KLVL:KLCH	110:99	52	1.54	17
	KLVH:KLCL	110:101	59	1.82	14
	KLVH:KLCH	110:99	52	1.85	10
	KLCL:KLCH	101:99	83	1.36	35
Dihydrofolate reductase	3DFR:4DFR	159:161	143	1.29	29
Lysozyme	1LZ1:LZHE	130:129	128	0.70	61
Plastocyanin/azurin	1PCY:1AZA	99:129	55	2.31	18
Papain/actinidin	PAP:2ACT	212:218	206	0.77	49

Proteins whose structure has been determined in different environments

				Reference
Trypsin inhibitor	58:58	56	0.40	Wlodawer *et al.*, 1984
Tuna cytochrome c	103:103	103	0.30	Takano and Dickerson, 1981
Azurin	129:129	127	0.37	Norris *et al.*, 1983
Rat protease	224:224	224	0.25	Anderson *et al.*, 1978
Deoxy human haemoglobin	287:287	287	0.30	Fermi *et al.*, 1984

[a]See Table I for abbreviations.

The relationship between the divergence of sequence and structure in the common cores of homologous proteins

The divergence of structure as measured by Δ is a simple function of the fractional sequence identity of the cores (Figure 2). A least squares fit to the data in Table II gives the relationship:

$$\Delta = 0.40\ e^{1.87H}$$

where Δ is measured in Å and H is the fraction of mutated residues. For the 32 pairs of homologous structures in Table II, the values of Δ predicted by this equation are within 20% of the observed values for 23 pairs and within 28% for the other nine.

The exponential form of the relationship arises because proteins accept mutations of surface residues more readily than mutations of buried residues. Closely related proteins differ primarily in surface residues, whereas distantly related proteins differ in both surface and buried residues (Table IV). The mutation of residues buried in the interior usually produces larger structural changes than the mutation of surface residues. Thus the tendency for changes in buried residues to lag behind surface changes results in an exponential relationship between sequential and structural change.

Conclusions

In a previous series of papers we have described the structural differences found in members of individual protein families (Lesk and Chothia, 1980, 1982. Chothia and Lesk, 1982, 1984). The differences in the common cores consist mainly of changes in the relative position and orientation of packed secondary structures and, in the case of β-sheets, some local changes in structure. We have shown here that the overall extent of these changes is directly related to the extent of the sequence differences.

These results imply that the degree of success to be expected in predicting the structure of a protein from its sequence using the known structure of an homologous protein, depends upon the extent of the sequence identity (Lesk and Chothia, 1986). A protein structure will provide a close general model for other proteins with which its sequence homology is >50%. If the homology drops to 20% there will be large structural differences that are at present impossible to predict.

However, the active sites of distantly related proteins can have very similar geometries (Lesk and Chothia, 1980; Chothia and Lesk, 1982; Read *et al.*, 1984). This is because of the coupling of the structural changes that has occurred during evolution (Lesk and Chothia, 1980). Thus the structure of the active site in a protein may provide a good model for those in related proteins even if the overall sequence homologies are low.

Table III. Cytochrome c family. Root mean square difference in the position of main chain atoms of residues in the conserved structural core, Δ

Protein pair[a]	Core determined for individual homologous pairs			Core common to four cytochrome c structures		
	Core size	Δ (Å)	Residue identity in core (%)	Core size	Δ (Å)	Residue identity in core (%)
3CYT:1CCR	103	0.62	59	48	0.38	65
3CYT:3C2C	99	1.13	37	48	0.91	48
1CCR:3C2C	101	1.47	41	48	1.01	56
3C2C:351C	56	1.50	36	48	1.39	35
3CYT:351C	57	1.65	25	48	1.56	31
1CCR:351C	58	1.86	24	48	1.66	27

[a]See Table I for abbreviations.

Table IV. The homology of buried and surface residues

Protein pair	Residue identity (%)		
	Buried residues[a]	Surface residues[a]	Overall
S. griseus proteases A and B	83	52	65
Human and hen egg white lysozyme	77	52	61
Tuna and rice embryo cytochrome c	77	50	59
Human haemoglobin α and *Chironomus* erythrocuorin	21	16	18
IgG Kol domains Vλ and $C\gamma_1$	31	11	17

[a]Buried residues are those with accessible surface areas ≤20 Å^2.

Acknowledgements

We thank Professor Sir David Phillips for the atomic co-ordinates of hen egg lysozyme, Professor J.Drenth for the atomic co-ordinates of papain, John Cresswell for the figure drawings and The Royal Society, National Science Foundation (PCM83-20171) and the National Institute of General Medical Science (GM25435) for support.

References

Almassy,R.J. and Dickerson,R.E. (1978) *Proc. Natl. Acad. Sci. USA*, **75**, 2674–2678.
Anderson,W.F., Matthews,B.W. and Woodbury,R.G. (1978) *Biochemistry*, **17**, 819.
Artymiuk,P.J. and Blake,C.C.F. (1981) *J. Mol. Biol.*, **152**, 737–762.
Baker,E.N. (1980) *J. Mol. Biol.*, **141**, 441–484.
Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.
Bhatia,G.E. (1981) Ph.D. Thesis, University of California at San Deigo.
Bolin,J.T., Filman,D.J., Matthews,D.A., Hamlin,R.C. and Kraut,J. (1982) *J. Biol. Chem.*, **257**, 13650–13662.
Chambers,J.L. and Stroud,R.M. (1979) *Acta Crystallogr.*, **35B**, 1861–1874.
Chothia,C. and Lesk,A.M. (1982) *J. Mol. Biol.*, **160**, 309–323.
Chothia,C. and Lesk,A.M. (1984) *J. Mol. Biol.*, **182**, 151–158.
Cohen,G.H., Silverton,E.W. and Davies,D.R. (1981) *J. Mol. Biol.*, **148**, 449–479.
Fermi,G., Perutz,M.F., Shaanan,B. and Fourme,R. (1984) *J. Mol. Biol.*, **175**, 159–174.
Furey,W., Wang,B.C., Yoo,C.S. and Sax,M. (1983) *J. Mol. Biol.*, **167**, 661–692.
Grace,D.E.P. (1979) D.Phil. Thesis, Oxford University.
Greer,J. (1981) *J. Mol. Biol.*, **153**, 1027–1042.
Guss,J.M. and Freeman,H.C. (1983) *J. Mol. Biol.*, **169**, 521–562.
Kamphuis,I.G., Drenth,J. and Baker,E.N. (1985) *J. Mol. Biol.*, **182**, 317–329.
Lesk,A.M. and Chothia,C. (1980) *J. Mol. Biol.*, **136**, 225–270.
Lesk,A.M. and Chothia,C. (1982) *J. Mol. Biol.*, **160**, 325–342.
Lesk,A.M. and Chothia,C. (1986) *Philos. Trans. R. Soc. Lond.*, **317**, 345–356.
Marquart,M., Deisenhofer,J., Huber,R. and Palm,W. (1980) *J. Mol. Biol.*, **141**, 369–391.
Matsuura,Y., Takano,T. and Dickerson,R.E. (1982) *J. Mol. Biol.*, **156**, 389–409.
Norris,G.E., Anderson,B.F. and Baker,E.N. (1983) *J. Mol. Biol.*, **165**, 501–521.
Ochi,H., Hata,Y., Tanaka,N., Kakudo,M., Sakurai,T., Aihara,S. and Morita,Y. (1983) *J. Mol. Biol.*, **166**, 407–418.
Perutz,M.F., Kendrew,J.C. and Watson,H.C. (1965) *J. Mol. Biol.*, **13**, 669–678.
Phillips,S.E.V. (1980) *J. Mol. Biol.*, **142**, 531–554.
Read,R.J., Fujinaga,M., Sielecki,A.R. and James,M.N.G. (1983) *Biochemistry*, **22**, 4420–4433.
Read,R.J., Brayer,G.D., Jurašek,L. and James,M.N.G. (1984) *Biochemistry*, **23**, 6570–6575.
Sielecki,A.R., Hendrickson,W.A., Broughton,C.G., Delbaere,L.T.J., Brayer, G.D. and James,M.N.G. (1979) *J. Mol. Biol.*, **134**, 781–804.
Steigemann,W. and Weber,E. (1979) *J. Mol. Biol.*, **127**, 309–388.
Takano,T. and Dickerson,R.E. (1981) *J. Mol. Biol.*, **153**, 95–115.
Wlodawer,A., Walter,J., Huber,R. and Sjölin,L. (1984) *J. Mol. Biol.*, **180**, 301–329.

Received on 27 January 1986

From genome sequences to protein function

Peer Bork, Christos Ouzounis and Chris Sander

EMBL, Heidelberg, Germany

A major goal of genome sequencing projects is the complete description of the function of all proteins. For most proteins sequenced in genome projects, an experimentally determined function is not available. Fortunately, evolutionary relationships can be exploited to predict the function of many other proteins from their amino acid sequence. The techniques for such predictions, sequence analysis by computational and database methods, are becoming increasingly sophisticated and are now an essential part of genome analysis.

Current Opinion in Structural Biology 1994, **4**:393–403

Introduction

Today, a sound PhD project in molecular biology often ends with the biochemical characterization and analysis of a cloned and subsequently sequenced gene. In the hope of additional functional insights, as well as interesting structural and evolutionary relationships, the sequence of the gene is usually further characterized by various computer methods. Often, the homologies found are extremely helpful for functional prediction. Today, the analysis of an average gene of about 1 kilobase (kb), including open reading frame (ORF) prediction, database searches, multiple alignment, pattern definition, and profile searches, might take several days or more, depending on the depth of the analysis. Soon genome projects will produce several hundred kilobases of raw sequences a day. Traditional sequence-analysis procedures cannot adequately handle such a high rate of production. The resulting dual challenges facing genome sequence analysis are both quantitative and qualitative: developing more efficient tools and increasing the scope of functional prediction. The work reviewed here represents the first steps on the way to meeting these challenges.

Tip of the iceberg: genomic sequence data today

In spite of the increased production of sequence data, today's databases contain only a small fraction of the genome's complete sequence. Many eukaryotic genome projects, including the Human Genome Initiative, are currently still assembling high-resolution physical maps [1,2], an essential prerequisite of systematic large-scale sequencing. So the real flood of genomic data is still to come.

Current output of genome sequencing projects

The most significant stretches of genomic sequences known to date are sizeable parts of relatively small genomes, such as the nematode *Caenorhabditis elegans*, the yeast *Saccharomyces cerevisiae*, and the bacterium *Escherichia coli* (Tables 1 and 2). In addition, there exist collections of numerous smaller chromosome segments from many different genomes, not yet assembled into long continuous stretches. The largest continuous sequence determined so far is a 2 Megabase (Mb) stretch of chromosome III of *C. elegans* [3], with another 5 Mb expected later in 1994 (R Durbin, personal communication). The next-largest pieces come from the European yeast genome project [4]: the complete sequences of five out of sixteen chromosomes will be available by the end of 1994 (Table 1), giving a total of 2.3 Mb, or 15% of the yeast genome. Bacterial genomes are next, with several large *E. coli* segments [5–7,8•,9,10] resulting in continuous pieces of up to 500 kb (Table 2). The total *E. coli* sequence data already covers more than 60% of the genome (P Rice, personal communication).

In less than 5 years, the genome projects of invertebrates such as the nematode [11], plants such as cress [12], vertebrates such as the buffer fish [13•], and mammals such as human [1], or mouse [2, 14] will probably generate hundreds of Mb of sequence data. Remarkably, the plans for genomic sequencing efforts appear to be on, or ahead of, schedule (Table 2) and are likely to meet the projected completion dates, provided the planned increases in funding are forthcoming. Completion of the *E. coli* and yeast sequencing efforts is expected by 1996. The *C. elegans* genome will be complete by the end of this decade (Table 2). In the human genome project, the initial estimates for completing sequencing were revised downward, from 2020 to 2010 and, most recently, to 2005 [1]. The methodologies for analyzing these data and, in particular, the

Abbreviations

3D—three-dimensional; **EST**—expressed sequence tag; **kb**—kilobase; **Mb**—megabase; **ORF**—open reading frame.

Table 1. Status of sequencing projects of model genomes.

Species	Size (Mb)	Sequenced (Mb)[a]	Sequenced (% of total)[a]	Year of completion[b]
Organelles/Viruses				
Mitochondria (various)	0.01–0.19		100%	1981
Chloroplasts (various)	0.12–0.16		100%	1986
Vaccinia virus	0.19		100%	1991
Prokaryotes				
Mycoplasmas	0.6–1.4	0.3	30%	(1996)
Mycobacterium leprae	2.8	0.1	4%	
Bacillus subtilis	4.2	2	40%	
Escherichia coli	4.7	3	60%	(1996)
Eukaryotes				
Yeast (*Saccharomyces cerevisiae*)	15	5	30%	(1998)
Chromosome III	0.31		100%	1992
Chromosome XI	0.66		100%	1994
Chromosome II	0.84		100%	1994
Cress (*Arabidopsis thaliana*)	100	1	1%	
Nematode (*Caenorhabditis elegans*)	90	3	3%	(1998)
Fruit fly (*Drosophila melanogaster*)	170	3	2%	
Mouse (*Mus musculus*)	3000	9	0.3%	
Human (*Homo sapiens*)	3600	19	0.6%	(2005)

[a]Note that the values are taken from the EMBL nucleotide sequence database; redundancies (only partly excluded) would lead to a decrease, while expressed sequence tags (not included) and data not yet released would lead to an increase in the figures given here.
[b]Estimated years of completion are in parentheses to indicate the approximate nature of the projections.

50 000–100 000 human genes will have to follow at the same pace!

The fast track to protein sequences: the EST approach

An attractive alternative to continuous sequencing is the random sequencing of reverse transcribed messenger RNA (cDNA) fragments, often called 'expressed sequence tags' (ESTs) [15]. This approach is 'quick and dirty' in that it is initially limited to single sequencing runs (no verification, no extension). However, it is ideal for very rapid identification of gene products as a first step toward elucidation of gene function, a major goal of genome projects. ESTs correspond to protein fragments of about 100 amino acids and can be used to obtain an expression profile for a particular organism or a tissue, to identify exons or to provide a glimpse into the molecular repertoire of various organisms. Already, a sizeable fraction of the ESTs sequenced have sequence similarity to proteins of known function. In 1991 only a few hundred ESTs from human brain were known [15]; these increased to more than 6000 human ESTs by 1993 [16••], and probably up to 100 000 sequences from several organisms will be available by the end of 1994. The rate of data production is so high that ESTs corresponding to most of the highly expressed human genes will probably approach saturation in 1995, many of them in private commercial efforts. Publicly available ESTs from various organisms are now in a specialized database, called dbEST [17].

Rising to the challenge: large scale sequence analysis

Coping with the analysis of rising amounts of sequence data has become a major scientific enterprise over the last few years, driven primarily by an abundance of information and its frequently incomplete digestion. Significant recent progress has been made in four main areas: first, development of databases; second, advances in computer networking; third, improvements in information access software; and fourth, improvements in algorithms and methods of sequence analysis by computer. We review each of these in turn.

Grow and link: development of databases

The usefulness of sequence databases, such as EMBL/GenBank (nucleic acids) and Swissprot/PIR (proteins), is currently limited by incomplete integration into a coherent whole, and by incomplete links to other biological databases, such as GDB (genome maps) and PDB (three-dimensional [3D] structures). A number of efforts are under way to provide more inter-database links (technical term: interoperability of databases), to add carefully annotated specialized databases, and to add value to existing databases by deriving additional information from them.

For example, numerous cross-pointers have been added to the Swissprot protein sequence database [18]. Given a protein sequence entry in this database,

Table 2. Principal sources of protein sequences.

Long genomic stretches	kb	ORFs[a]	References[b]
Nematode chromosome III	2000	n.d.	[3]
Yeast chromosome II	840	387	Feldmann *et al.*
Yeast chromosome XI	666	331	Dujon *et al.*
Escherichia coli[c]	500	n.d.	[8•]
Yeast chromosome III	320	170	[4,49•]
Cytomegalovirus	229	190	
Vaccinia virus	191	74	
Liverwort mitochondrion	187	74	
Tobacco chloroplast	156	109	
Bacillus subtilis	97	92	[54]
Fruit fly homeobox loci	80	n.d.	
Bacteriophage λ	49	63	
Mycobacterium leprae	37	12	[55]
Contig/EST collections	**kb**	**pieces**	
Mycoplasma contigs	350	650	[56]; Gillevet *et al.*
Caenorhabditis elegans (nematode) EST	n.d.	4699	[17,57,58]
Human brain EST	n.d.	6000	[16••]
Arabidopsis thaliana (cress) EST	n.d.	4512	[17]
Oryza sativa (rice) EST	n.d.	4231	[17]
Databases[d]	**kb or kaa[a]**	**seqs[a]**	
Swissprot[e](kaa)	11 500	33 300	
EMBL[e]without ESTs (kb)	155 000	155 000	
dbESTs [kb]	~10000	31 800	
PDB [kaa]	444	2200	

[a]Abbreviations: kaa, 1000 amino acids; kb, kilobases; n.d., not determined, or not available to the authors; ORFS, number of open reading frames (predicted proteins); seqs, number of sequence entries. [b]Selected references only are given; unnumbered references are personal communications. [c]There are four neighbouring, slightly overlapping *E. coli* segments in the EMBL nucleotide sequence database — the longest one covers about 176 kb. [d]Redundancy within and among sequence databases slightly reduces the numbers of bases/genes given; PDB = Protein Data Bank of three-dimensional structures. [e]Similar numbers for PIR (proteins) and GenBank (DNA).

cross-references give information, if available and applicable, on the nucleic acid sequence of its gene, the genomic map location of the gene, the enzyme commission functional classification, the three-dimensional structure, known homologs, known characteristic sequence patterns (PROSITE [19]), and so on. An example of added-value databases is the database of sequence–structure alignments (HSSP) that provides multiple sequence alignments, position-specific information on the degree of conservation, and sequence search profiles [20]. Examples of specialized databases, designed to address particular end-user requirements and carefully maintained by curators, are the databases of ESTs [17], *E. coli* databases [21,22], the worm (nematode *C. elegans*) database (see above), and FLYBASE (*Drosophila*) [23]. There are many more, too numerous to list. The urgent need for cross-references and interoperability is illustrated by the fact that a directory of molecular biology databases is a database in itself, e.g. LiMB [24].

Dial-up information: wide-area computer networks

Databases in themselves are of limited use unless they are easily accessible. Computer networks and information-retrieval software provide such access. Recently, both the physical and logical infrastructure of national and international computer networks and network software have improved considerably. As a result, use of wide-area information resources has become an important activity for sequence analysis experts. These resources allow the search, identification, and subsequent transfer of large amounts of publicly available material, including software and data in various forms. Information typically travels on Internet, the precursor of the planned 'information superhighways'. National and international laboratories provide various servers for automated sequence analysis [25]. Database updates can be performed using file transfer protocols (ftp). Resource browsers such as XMosaic™, developed at the US National Center for Supercomputer Applications (NCSA) in Illinois, are valuable software

tools for navigating networks in search of data and information.

As more and more institutions offer data and services, the identification of relevant information resources becomes increasingly difficult. What is needed are client–server systems and information service brokers. In such systems, the user specifies the tasks using local (client) software and sends out the request to a remote system (server) which then distributes the required actions over the appropriate and available resources (for example, the 'Hassle' prototype [26] being tested on EMBnet, the European Molecular Biology Network). The key point is that copies of all data, software and information are not, and cannot, be kept locally without considerable effort, and therefore convenient access to network information services is an essential element for efficient sequence analysis work.

Access to integrated information: software tools

An interesting development in the direction of intelligent retrieval systems for genome data is AceDB (**A** ***C***. ***e****legans* **d**atabase, by Richard Durbin and Jean Thierry-Mieg), an integrated system developed for the nematode sequencing project, now also in use by other genome projects. AceDB incorporates many programs that handle and analyze raw DNA and other sequence data, as well as map data, and contains a convenient graphical user interface with hyperlinks for data browsing. Other systems are under development, each with slightly different orientation, at the US National Center for Biotechnology Information (NCBI), the Genome Database Center at John Hopkins Medical School (GDB), EMBL, the German Cancer Research Center (DKFZ) and elsewhere. An excellent example of future client–server access to information across international boundaries is the network version of the Entrez software distributed by the NCBI that gives access to Medline literature citations relevant to sequence databases. If integrated access to diverse information becomes generally available, the efficiency of sequence analysis work will be much higher. To make effective use of all the information, genome projects will have to develop and apply techniques for information services in parallel with sequencing technology.

Progress report: methods of sequence analysis

Common sequence analysis practiced by the casual user, using standard programs, tends to miss a significant fraction of the functional information in protein sequences. It is therefore important to see what can be achieved with new and sophisticated methods. The list of basic procedures (selection in Table 3) is gradually increasing as new algorithms are invented and old ones improved (reviewed in [27]). We review here some of the most interesting programs that have had a practical impact on the field.

Identification of open reading frames

The accuracy of ORF prediction has been improved [28,29], relative to widely used programs such as Genmark [30], Genefinder [31], Grail [32] and the Staden package [33]. Frameshift detection and detection of other errors are an important technical issue, affecting the quality of derived amino acid sequences [34,35]. Routine use will indicate which of these new methods have noticeable practical advantages.

Analysis of amino acid composition bias

Once a putative protein sequence is available, a number of analysis methods can be applied. The best known, and most powerful, are sequence database alignment searches. However, an assessment of the significance of particular 'hits' (match of query with target sequence) depends strongly on the structural and functional class of the protein coded by that sequence (globular, filamentous, transmembrane, etc.). In practice, the problem is twofold. First, standard measures of sequence similarity require different cutoffs depending on the amino acid composition of the sequences being compared. Second, larger proteins have an inhomogeneous amino acid composition, i.e. a distinctly different composition in different regions (e.g. hydrophobic or charged stretches or 'domains'). Thus, before doing alignment database searches, it is useful to first determine the composition bias, i.e. deviation from the typical composition of globular proteins or deviation from the average composition in the protein sequence database. An approximate classification of different types of composition bias is as follows:

Coiled-coil arrangements. Whereas the detection of such regions using the program of Lupas *et al.* [36] appears to be fairly accurate, no reliable distinction between two-stranded and three-stranded, or between parallel and antiparallel, coiled-coil regions is possible yet.

Transmembrane regions. There are many programs for the prediction of transmembrane segments and signal peptides. Most are based on the simple notion of detecting runs of hydrophobic residues. We are not aware of any recent substantial improvement in accuracy.

Other low-entropy (or low-complexity) regions. A particular amino acid composition may be the result of functional selective pressure, e.g. a run of positively charged residues involved in non-specific protein–nucleic acid interaction. Recently, progress has been made in identifying heavily biased regions such as small repeats, long charged clusters, or regions rich in one, or a few, particular amino acids generally atypical of globular proteins (Wootton, this issue, pp 413–421). Complementing the work of Karlin and associates [37], the programs 'Seg' [38•] and 'Xnu' [39•] fill an urgent need in this area. These methods identify and mask composition-biased regions in the query sequence. The surrounding sequence and the biased regions can then be processed separately.

Table 3. From genome sequences to protein function and structure.

Steps	Problem	References[a]
Contig assembly	Detection of DNA overlaps	
Error correction	Identification of frameshifts, etc.	[34,35]
Open reading frame prediction	Identification of putative genes	[28,29]
Masking	Exclusion of regions with amino acid composition bias	[38•,39•]
Coiled-coil detection	Recognition of coiled-coil areas	[36]
Hydrophobicity analysis	Detection of transmembrane and signal segments	
Database homology search	Detection of sequence similarities	[41•,43••]
Multiple alignment and tree construction	Definition and analysis of protein families; definition of profiles	[41•,44,59]
Database profile or pattern search	Detection of distant relationships	
Self-alignment	Detection of internal repeats	
Secondary structure prediction	Prediction of helices, strands, loops, and surface/interior	[45]
Three-dimensional modelling	Construction of detailed atomic models based on homology	

[a]Only references to some recent developments are included.

Derivation of amino acid comparison matrices

All alignment search methods use scoring matrices that assign similarity values for any pair of the 20 amino acids. A new scoring matrix, 'Blosum', was derived from multiple-sequence alignments in a database of conserved regions ('Blocks' [40]). Tests indicate that use of the Blosum matrix leads to improved performance in homology detection [41•] and its use is therefore increasing.

Database searches

A search for sequence similarities in a protein sequence database is the most useful and most widely used method of functional prediction. The limitations arise from the fact that derived amino acid sequences enter the databases only after some delay (due to processing) or not at all (due to errors in nucleic acid sequences, or in their interpretation). It is therefore useful, albeit costly, to search six-frame translations of nucleic acid sequences. One tool in the 'Blast' series [42] of fast search programs, TBlastN, does this very effectively. On occasion, this mode of search reveals homologies with very recently sequenced genes or with adjacent sequences in two different reading frames, usually evidence of a frameshift, possibly as a result of a sequencing error. The reverse operation, searching the protein sequence database with the six-frame translations of a newly sequenced gene using BlastX [43••], is extremely useful for characterizing protein coding regions in raw nucleic acid sequence data, thus complementing ORF prediction programs.

If several putative homologs are detected in a database scan, multiple-sequence alignment can reveal common functional motifs. A new method for the automatic detection of motifs in a set of sequences [44] now offers an alternative to standard programs. The use of profiles derived from multiple-sequence alignments also lead to improved secondary structure prediction [45].

Despite the progress in basic analysis techniques, the interpretation of apparently significant sequence similarities in functional terms is still an underdeveloped area. More sophisticated methods for the analysis of a set of sequences in a family are needed for a more accurate homology-based prediction of protein function.

The key goal: prediction of protein function

When all the analytical procedures (Table 3) have been applied to a newly determined protein sequence of unknown function, one faces the difficult task of assigning a putative function based on evidence from sequence similarities, pattern detection, and so on. This part of the analysis process cannot easily be embodied in an algorithm and therefore is the least automated. Typically, an attempt is made to interpret sequence similarity by transferring the functional information about one protein to the homologous relative. The problem is that prediction of function by analogy can be very precise and complete, or it can be very fuzzy and incomplete, depending on the data at hand and on the precise nature of the similarity.

The simplest cases are those in which there is strong sequence similarity to, for example, an enzyme of known function. The new protein is then predicted to have the same function as its homolog. But even in simple cases caution is needed. Sequence variation can have diverse consequences. Even highly similar proteins might have completely different functions. Striking examples are the eye lens crystallins, structural proteins which apparently evolved recently from metabolic enzymes [46]. Furthermore, isoenzymes fine-tune the metabolism by varying only slightly in substrate affinity; the substrate specificity of clearly homologous proteins can be changed completely, as observed in the different families of sugar kinases [47]. An unambiguously homologous protein may be the equivalent gene in another species, indicating direct lineage (ortholog) or the homology may merely imply descent after gene duplication (paralog). In the latter case, one has to be particularly cautious with functional predictions.

In other cases, structural or functional similarity may occur for only a small part of a larger given protein (a domain), e.g. from the presence of a zinc finger motif. Although this may suggest that the protein binds DNA or RNA, little else can be concluded about the function of the protein of which this domain is a part. Finally, only a few proteins, mainly enzymes, are sufficiently well characterized so that molecular details of catalysis as well as higher levels of regulation and physiological roles can be described.

These problem cases illustrate the need for more precise knowledge about functional variation in evolution. Which functional changes occur as a result of certain types and certain amounts of sequence change? More biochemical and genetic characterizations of sets of homologous proteins (or of engineered mutants of natural proteins) are needed before more precise and quantitative rules for function prediction by homology can be formulated.

Yield: how much functional information from homology ?

Considering the limitations described above, homology- and analogy-based predictions have proved to be extremely powerful in exploring genomic information. When comparing the output of several large-scale sequencing projects (Table 2), an identification of partial function has been possible for 40–65% of all predicted proteins (Fig. 1), with this figure showing an increasing tendency [16••,48,49•]. Remarkably, the number of tentative functional identifications by homology exceeds by far the number of functional determinations by direct experiment (Table 4). The first complete chromosome sequences have led to the rather sudden realization that we already know a considerable fraction of all protein functions. In yeast chromosome III, the most carefully analyzed eukaryotic chromosome to date, the rate of tentative functional assignment by homology exceeds 50% (Table 4).

Extrapolating into the future, we can expect a rapid rise in the probability of homolog detection when comparing a new sequence with all known sequences (Green, this issue, pp 404–412). This is particularly true if the sequencing of human ESTs approaches saturation in 1995, as has been predicted (M Adams, personal communication). However, the rate of functional prediction by homology will not rise nearly as fast. This is simply because on the one hand new sequences enter the databases at a rapid rate, while on the other hand most of these new sequences come without primary, i.e. experimental, functional information. So the gap between the total number of known protein families and those with identified function (Table 4) will increase considerably in the near future, before it ultimately decreases near the saturation limit.

These observations have two implications for the allocation of resources in genome projects. To maximize functional information, it is probably advisable to complement sequencing effort with two key activities. One focus is a concerted program for the experimental determination of new protein functions, in order to increase the store of primary functional information on which all derived functional identifications depend. The other focus is a further improvement in the reliability of homology detection by sequence data analysis, in order to make maximum use of the experimental information. Both activities are essential and interdependent. On the one hand, even a single experimentally determined function can be immediately carried over, at least in part, to all of its sequence relatives and represents a sizeable net gain in information, if it is of a new type. On the other hand, even a single percentage

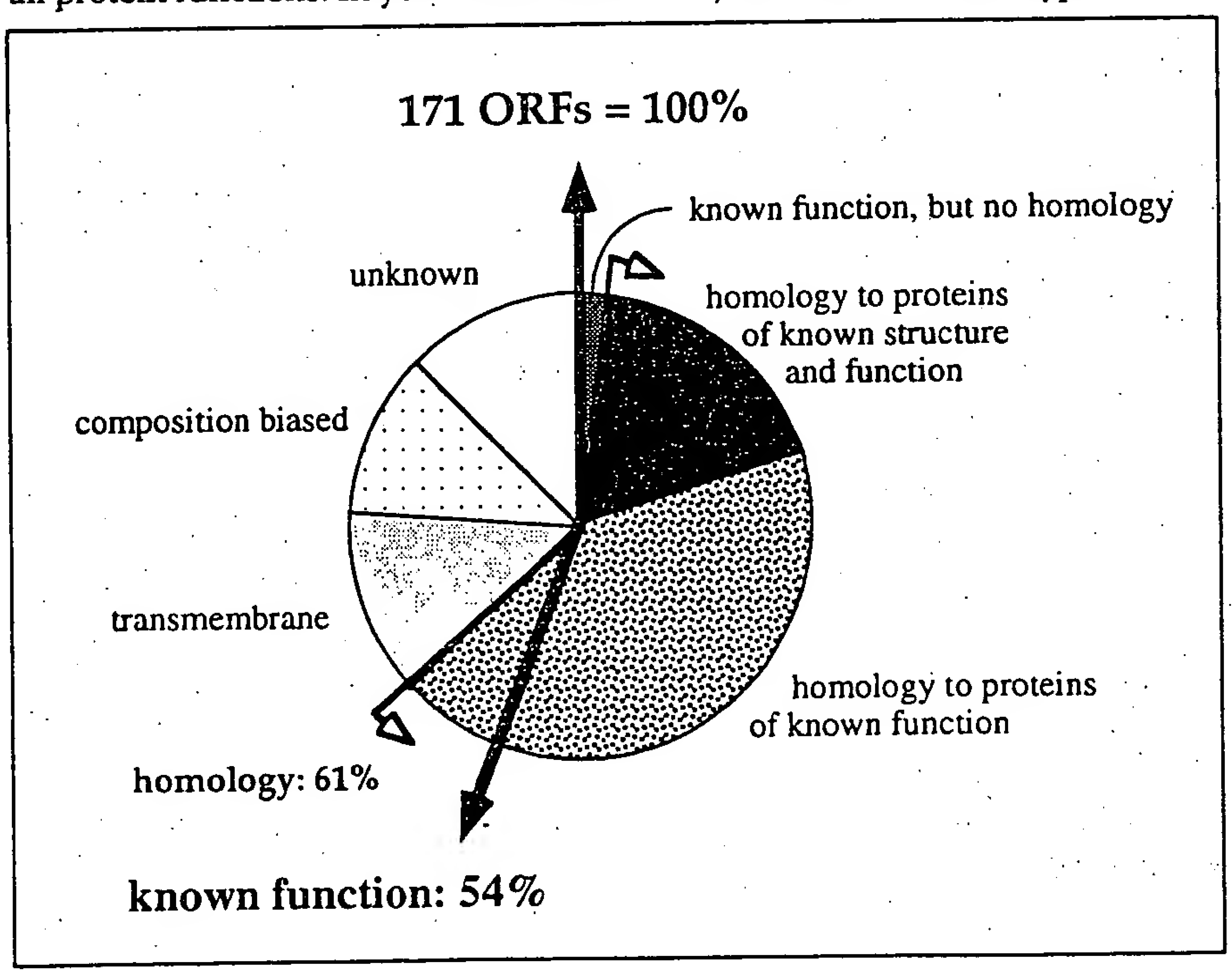


Fig. 1. Information content in the proteins of yeast chromosome III [49•]. There is a 10% gap between all protein families ('homology') and those of known function, termed here the function–homology gap (see text and Table 4). Numerically: 10% = 61% (homology) − 54% (known function) + 3% (known function, no homology).

Table 4. Identification of protein function by experiment and by homology, showing an increase in the function–homology gap using *Saccharomyces cerevisiae* chromosome III as an example.

	Yeast chromosome III (as of 1/1993)	Yeast chromosome III (as of 1/1994)	Anticipated increase or decrease[a]	Speculative estimate for yeast (1995/6)
Function[b]				
Function by experiment	11%	14%	+	17%
Function only by homology	31%	40%	++	50%
No function yet	58%	46%	––	33%
Homology[c]				
Member of sequence family	38%	61%	+++	80%
No family yet	62%	39%	–––	20%
Function–homology gap[d]	**1%**	**10%**		**15%**

[a]Increase: slow (+), intermediate (++) and rapid (+++); same for decrease (––), (–––). [b]Known full or partial function is given, identified by direct experiment or by detection of homology to a protein of known function (data from [4,49•,50]). [c]Experimental function represents an estimate, as there is some ambiguity in the definition of function; e.g. 'temperature sensitive lethal' was not counted as known function, while 'DNA repair protein' was counted. [c]Homology represents a significant sequence similarity to at least one other protein in the sequence databases, thus defining a protein family of two or more sequences, with or without known function. [d]'Gap' represents the fraction of all known protein sequences which have at least one homolog (belong to a family), but for which the function is not yet known, and was estimated using the numbers in the table and the number of proteins with known function that have no known homolog; the reason for the increasing gap is the determination of numerous new protein sequences without known function.

point improvement in functional prediction (currently at about 50%) translates to a large absolute number of identified protein functions, with immediate savings in experimental effort.

Molecular detail: implied three-dimensional structures

Detection of significant sequence similarity to a protein of known 3D structure immediately implies prediction of the 3D structure of the new protein by homology. The prediction leads to detailed arguments about the mechanism of protein action and about the role of particular residues. Often, very conserved residues distributed along the protein chain are in spatial proximity in 3D, explaining previously puzzling conservation patterns and suggesting detailed experiments. It is therefore interesting that as many as 19% of all yeast chromosome III sequences are significantly homologous to a known 3D structure (Fig. 1). For these proteins, approximate atomic models can be built, inspected using 3D graphics, and used as a basis for planning experiments. However, coverage is incomplete for certain types of protein structures, e.g. membrane proteins of which only a handful of 3D structures are known.

As the rate of 3D structure determination of proteins now exceeds one structure a day, the rate of 3D prediction by homology will probably reach the 20–25% level in the not so distant future. It already exceeds 24% for the Swissprot database (R Schneider, personal communication). This number comes from a systematic all-against-all alignment of the sequences in the Protein Data Bank (known structures) with the 32 000 sequences in Swissprot [20], assessing homology by the application of a threshold for structural similarity. It is remarkable that we already have the opportunity to identify the approximate tertiary structure of one in four newly sequenced proteins.

Feedback: checking predictions by experiment

The experimental verification of predictions by homology is very slow. The higher the belief in the validity of computer-based sequence analysis, the lower the incentive to perform control experiments. However, there is a small but non-negligible number of cases where prediction entices experiment, for a variety of reasons. One case is the surprising occurrence of a certain functional type in another organism.

An example of this is the discovery of a certain type of polymerase in yeast. The homology and predicted functional analogy between yeast ORF YCR14C and mammalian DNA polymerase-β [48,50] was experimentally verified by *in vitro* tests on the expressed protein [51], now named yeast DNA polymerase IV (*pol IV* gene) — an example of a successful prediction based on multiple alignment and sequence pattern analysis [50,52]. This now opens the way to biochemical and genetic studies in yeast, not just mammals, to understand the presumed role of the short gap-filling activity of β-polymerases in DNA repair.

Improvements: what to expect

The main gain in the level of function prediction over the last 2 years came primarily from three sources. First, protein-sequence databases have improved in quantity (more sequences, more frequent updates) as well as in quality (better annotation and more cross-pointers). Second, multiple-sequence alignment methods have improved, especially profile and pattern definition for protein families and their detection in new sequences. Third, diverse information resources are more conveniently accessible over the Internet, lowering the threshold for careful analysis, and controls and cross-checks for borderline cases. Full use of these improvements typically requires expertise.

```
MBHA_MYXXA    34 VALDIKSDDGGKTLKGTMT  myxobacterial hemagglutinin
MBHA_MYXXA   101 VAVSIKSNDGGKTLTGTTT  (Myxococcus xanthus)
MBHA_MYXXA   168 INVDAKSNDGGKTLSGTMT
MBHA_MYXXA   235 VALNVASSDGGKTLAGTMI
NANH_BACFR     1 DVGLSRSTDGGKTWEKMRL  sialidase (EC 3.2.1.18)
NANH_BACFR    80 QLVLAKSTDDGKTWSAPIN  (Bacteroides fragilis)
NANH_BACFR   140 NAGIMYSKDGGKNWKMHNY
NANH_BACFR   247 NTTIKISLDGGVTWSPEHQ
NANH_CLOSE   426 DTGIKRSTDGGVTWDEGKI  sialidase precursor (EC 3.2.1.18)
NANH_CLOSE   559 FLSLIYSDDDGQTWSDPID  (Clostridium septicum)
NANH_CLOSE   623 SSAVIYSDDNGATWNIGET
NANH_CLOSE   696 RVRIATSFDGGATWEDDVV
NANH_SALTY    66 DTAAARSTDGGKTWNKKIA  sialidase (EC 3.2.1.18)
NANH_SALTY   140 DLVLYKSTDDGVTFSKVET  (Salmonella typhimurium)
NANH_SALTY   205 NTSFIYSTD-GITWSLPSG
NANH_SALTY   251 LRRSFETKDFGKTWTEFPP
TCNA_TRYCR    19 DTVAKYSVDDGETWETQIA  sialidase (EC 3.2.1.18) SA85-1.2 -
TCNA_TRYCR   130 TPEVTKSTAGGKITASIKW  major surface antigen
TCNA_TRYCR   159 FSKIFYSEDDGKTWKFGKG  (Trypanosoma cruzi)
TCNA_TRYCR   205 RRLVYESSDMEKPWVEAVG
PEP1_YEAST   411 KGVTKISVDNGLTWTMLKV  PEP1 on yeast chromosome II
PEP1_YEAST   483 DQRTFISRDGGLTWKLAFD  (Saccharomyces cerevisiae)
PEP1_YEAST   529 QSEFYYSLDQGKTWTEYQL
PEP1_YEAST   581 TTNFIYAIDFSTAFNDKTC
YCZ0_YEAST    63 LSEIFISDSQGLKFSPIPF  YCR100C on yeast chromosome III
YCZ0_YEAST   120 GGETKISVDNGLTWSNLKV  (Saccharomyces cerevisiae)
YCZ0_YEAST   192 DRKTFISRDGGLTWRVAHN
YCZ0_YEAST   238 QSKLYFSLDQGRTWNQYEL
motif:               hhhS D G TW        (h - hydrophobic)
```

Fig. 2. Sialidase sequence motif in yeast proteins [49•]. A typical example of short and weak motifs that are not detectable by standard homology searches. The motifs are, however, detectable by profile and pattern searches. The significance of the sialidase motif is supported by its multiple occurrence with an average spacing of about 50 amino acids. As a result, a sialidase activity is predicted for the yeast proteins. As a 3D structure of one bacterial sialidase is now known (Fig. 3), structural, as well as functional, information can be inferred for the yeast proteins.

An example of a result obtained with more sophisticated methods in the context of large-scale genome analysis is the identification of sialidases in yeast chromosomes III and II. During the analysis of chromosome III [49•,50] the putative ORF YCR100 initially did not match any database protein apart from PEP1, a functionally uncharacterized protein from yeast chromosome II. However, the conservation profile between both proteins revealed four short, but conserved, internal repeats. Using this information in pattern search methods [52], the investigators detected subtle similarities to conserved repeats in bacterial and protozoan sialidases (Fig. 2). Very recently, the three-dimensional structure of one of these sialidases was determined [53] (Fig. 3), and indeed, it contains internally repeated β-sheets forming a superbarrel or propeller fold, fully consistent with the sequence repeat; the most conserved residues (Fig. 2) are located in equivalent positions in the respective sheets (Fig. 3). Thus, based on a short signature motif (too short and weak to be detected by conventional homology search programs) a rather precise functional and structural prediction was made. This example emphasizes the need to incorporate such methods into standard analysis procedures and illustrates the potential gain.

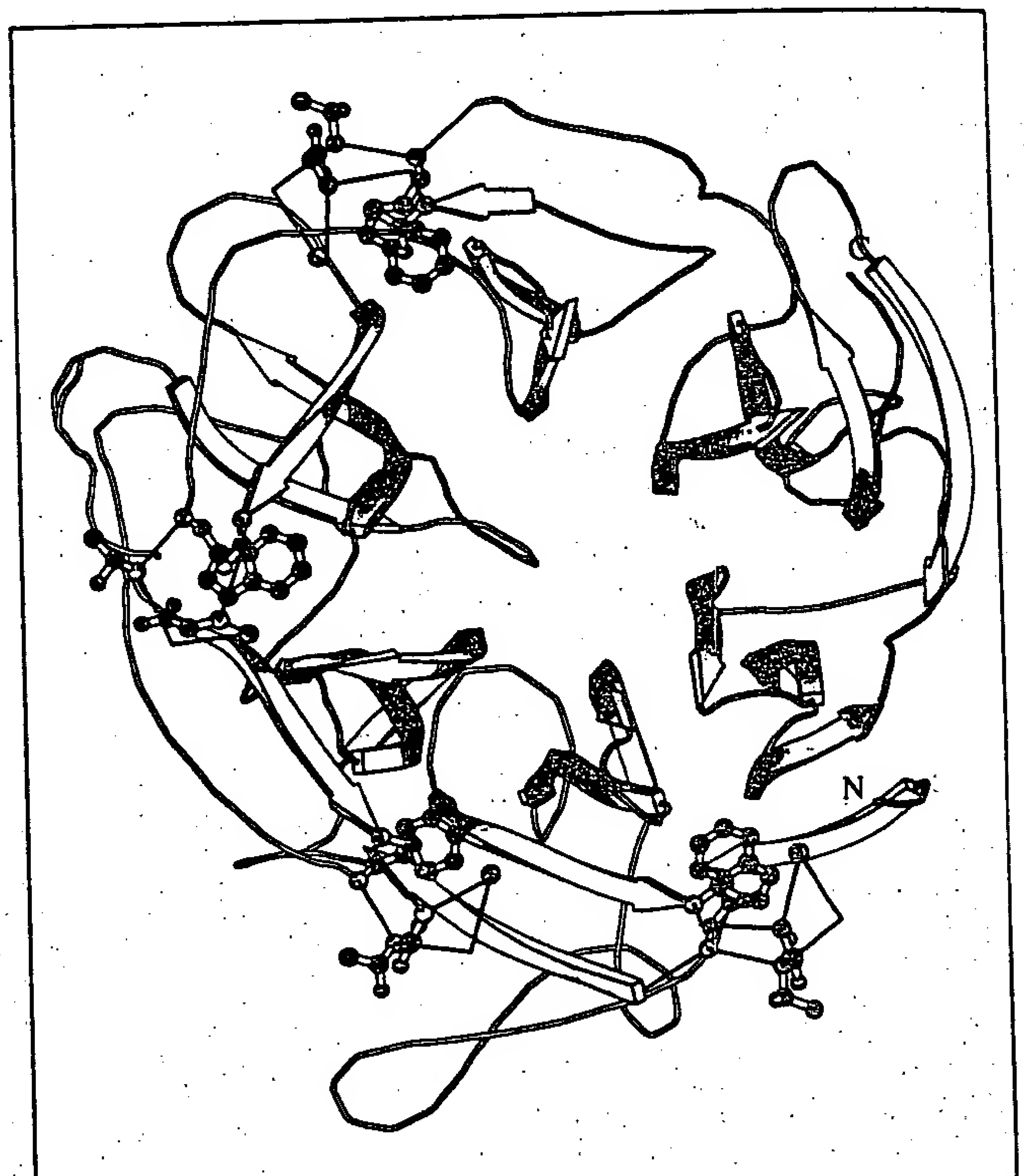


Fig. 3. Ribbon plot of the sialidase from *Salmonella typhimurium* [53]. It is an example of the so-called β-propeller fold in which six four-stranded β-sheets form a superbarrel. The four sequence motifs are located in equivalent positions in four of the six β-sheets. In the remaining two sheets, the corresponding motifs are not detectable (Fig. 2).

In summary, for the immediate future the most pressing and promising needs of genome sequence analysis are manifold. First, further refinement of pattern and profile searches. Second, automation of the analysis process, especially for sequence families. Third, improved data support by direct access to specialized and bibliographic databases. Fourth, earlier public accessibility of data from major sequencing projects. Fifth, training of analyzers in advanced methods. Sixth, the development of a more refined classification of protein function, and finally, a better

understanding of the effect of sequence changes on protein function.

Conclusion

In a few years' time, the complete sequences of several entire genomes will become known, resulting in a series of historical achievements: *E. coli*, mycoplasma(s), yeast, nematode, are proceeding apace (Table 1). Genomes from organisms of all taxonomic ranks will follow, including *'Homo ignorans'*. The analysis of these data will require a high degree of automation in sequence analysis without sacrificing the sensitivity of present methods for the detection of distant sequence similarities. Where sequence analysis has no answers, experimental technologies will be essential for the genetic and biochemical characterization and the physical identification of completely new types of proteins. In addition, many experiments will be stimulated by the detailed predictions of sequence analysis. The database — or is it knowledge base? — of all proteins will gradually be completed, including 3D structures and diverse functional knowledge.

Imagine that at some time almost all proteins of a particular organism will be known — sequence, 3D structure and function — and stored in the World-Prot database. The obvious question that then arises is whether the molecular repertoire of an organism is sufficient to characterize its physiological and evolutionary behavior. The obvious answer is that it is not, and that biological experiments and theories at higher, less microscopic, levels are needed to complement the atomic information available from genome-sequencing projects.

Will a graduate student today, trained in sequencing and in sequence analysis, end his career like a zoologist in the beginning of this century, hunting the last unclassified butterflies in Madagascar? In other words, will experimental and computational sequence analysis be transformed from a skilful scientific endeavor to an activity of lesser scientific interest? Or, will it provide the ultimate answers to the grand questions of biological science, about structure and function, development and evolution? The truth lies somewhere in between. The functional classification of all proteins will be an excellent intermediate goal for that graduate student, but also an excellent point of departure for addressing the real questions of human health, the environment, and the future evolution of life.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:
- • of special interest
- •• of outstanding interest.

1. Collins F, Galas D: A new Five-Year Plan for the US Human Genome Project. *Science* 1993, 262:43–46.

2. Copeland NG, Jenkins NA, Gilbert DJ, Eppig JT, Maltais LJ, Miller JC, Dietrich WF, Weaver A, Lincoln SE, Steen RG, *et al.*: A Genetic Linkage Map of the Mouse: Current Applications and Future Prospects. *Science* 1993, 262:57–66.

3. Wilson R, Ainscough R, Anderson K, Baynes C, Berks M, Bonfield J, Burton J, Connell M, Copsey T, Cooper J, *et al.*: The *C. elegans* Genome Project: Contiguous Nucleotide Sequence of Over Two Megabases from Chromosome III. *Nature* 1994, 368:32–38

4. Oliver SG, van der Aart QJM, Agostoni-Carbone ML, Aigle M, Alberghina L, Alexandraki D, Antoine G, Anwar R, Ballesta JPG, Benit P, *et al.*: The Complete DNA Sequence of Yeast Chromosome III. *Nature* 1992, 356:38–46.

5. Wang MX, Church GM: A Whole Genome Approach to *in vivo* DNA-Protein Interactions in *E. coli*. *Nature* 1992, 360:606–610.

6. Plunkett G III, Burland V, Daniels DL, Blattner FR: Analysis of the *Escherichia coli* Genome. III. DNA Sequence of the Region from 87.2 to 89.2 Minutes. *Nucleic Acids Res* 1993, 21:3391–3398.

7. Burland V, Plunkett G III, Daniels DL, Blattner FR: DNA Sequence and Analysis of 136 kilobases of the *Escherichia coli* Genome: Organizational Symmetry Around the Origin of Replication. *Genomics* 1993, 16:551–561.

8. • Blattner FR, Burland V, Plunkett G III, Sofia HJ, Daniels DL: Analysis of the *Escherichia coli* Genome. IV. DNA Sequence of the Region from 89.2 to 92.9 Minutes. *Nucleic Acids Res* 1993, 21:5408–5417.

This is the last report in a series of papers that describe long segments of the *E. coli* genome. With this report, the data from this model organism can be assembled to a continuous segment of about 500 kb.

9. Yura T, Mori H, Nagai H, Nagata T, Ishihama A, Fujita N, Isono K, Mizobuchi K, Nakata A: Systematic Sequencing of the *Escherichia coli* Genome: Analysis of the 0–2.4 Min Region. *Nucleic Acids Res* 1992, 20:3305–3308.

10. Daniels DL, Plunkett III G, Burland V, Blattner FR: Analysis of the *Escherichia coli* Genome: DNA Sequence of the Region from 84.5 to 86.5 Minutes. *Science* 1992, 257:771–778.

11. Sulston J, Du Z, Thomas K, Wilson R, Hillier L, Staden R, Halloran N, Green P, Thierry-Mieg J, Qiu L, *et al.*: The *C. elegans* Genome Sequence Project: a Beginning. *Nature* 1992, 356:37–41.

12. Reiter RS, Williams JG, Feldmann KA, Rafalski JA, Tingey SV, Scolnik PA: Global and Local Genome Mapping in *Arabidopsis thaliana* by Using Recombinant Inbred Lines and Random Amplified Polymorphic DNAs. *Proc Natl Acad Sci USA* 1992, 89:1477–1481.

13. Brenner S, Elgar G, Sandford R, Macrae A, Venkatesh
• B, Aparicio S: Characterization of the Pufferfish (*Fugu*) Genome as a Compact Model Vertebrate Genome. *Nature* 1993, 366:265–268.

The authors identify the *Fugu* genome as a model vertebrate genome for sequencing projects, due to its small size (only four times larger than *C. elegans*). Evidence is presented that this particular genome is 90% unique, and devoid of repetitive sequences.

14. Chapman VM, Copeland NG, Costantini FD, Dove WF, Nadeau JH, Reeves RH, Rossant J, Smithies O, Woychik RP: A Plan for the Mouse Genome Project. *Mammalian Genome* 1993, 4:293–300.

15. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, *et al.*: Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project. *Science* 1991, 25:1651–1656.

16. Adams MD, Kerlavage AR, Fields C, Venter JC: 3,400 New
•• Expressed Sequence Tags Identify Diversity of Transcripts in Human Brain. *Nature Genet* 1993, 4:256–267.

The authors provide a very interesting analysis of the ever-increasing EST data. A profile of the transcriptional activity of human brain is presented, based on sequence similarity searches. Cytoskeletal and other structural proteins seem to be the most abundant.

17. Boguski MS, Lowe TMJ, Tolstoshev CM: dbEST Database for Expressed Sequence Tags. *Nature Genet* 1993, 4:332–333.

18. Bairoch A, Boeckmann B: The SWISS-PROT Protein Sequence Data Bank, Recent Developments. *Nucleic Acids Res* 1993, 21:3093–3096.

19. Bairoch A: The PROSITE Dictionary of Sites and Patterns in Proteins, its Current Status. *Nucleic Acids Res* 1993, 21:3097–3103.

20. Sander C, Schneider R: The HSSP Data Base of Protein Sequence-Structure Alignments. *Nucleic Acids Res* 1993, 21:3105–3109.

21. Kröger M, Wahl R, Rice P: Compilation of DNA Sequences of *Escherichia coli* (Update 1993). *Nucleic Acids Res* 1993, 21:2973–3000.

22. Rudd KE: Maps, Genes, Sequences, and Computers: An *Escherichia coli* Case Study. *Am Soc Med News* 1993, 59:335–441.

23. FLYBASE: The Drosophila Genetic Database, 1993. Available from ftp.bio.indiana.edu network server.

24. Lawton JR, Martinez FA, Burks C: Overview of the LiMB Database. *Nucleic Acids Res* 1989, 17:5885–5899.

25. Henikoff S: Sequence Analysis by Electronic Mail Server. *Trends Biochem Sci* 1993, 18:267–268.

26. Doelz R: Hassle — A Tool to Access Sequence Databases Remotely. *Comput Appl Biosci* 1994, 10:31–34

27. Doolittle RF: Protein Sequence Comparisons: Searching Databases and Aligning Sequences. *Curr Opin Biotechnol* 1994, 5:24–28.

28. Guigo R, Knudson S, Drake N, Smith T: Prediction of Gene Structure. *J Mol Biol* 1992, 226:141–157.

29. Farber R, Lapedes A, Sirotkin K: Determination of Eukaryotic Protein Coding Regions Using Neural Networks and Information Theory. *J Mol Biol* 1992, 226:471–479.

30. Borodovsky M, McIninch J: Recognition of Genes in DNA Sequences with Ambiguities. *Biosystems* 1993, 30:161–171.

31. Green P, Hillier L: Genefinder Software, Unpublished. 1993, Department of Genetics, University of Washington, Missouri.

32. Uberbacher E, Mural R: Locating Protein-Coding Regions in Human DNA Sequences by a Multiple Sensor-Neural Network Approach. *Proc Natl Acad Sci USA* 1991, 88:11261–11265.

33. Staden R: Finding Protein Coding Regions in Genomic Sequences. *Methods Enzymol* 1990, 183:163–180.

34. Posfai J, Roberts RJ: Finding Errors in DNA Sequences. *Proc Natl Acad Sci USA* 1992, 89:4698–4702.

35. Claverie J-M: Detecting Frameshifts by Amino Acid Comparison. *J Mol Biol* 1993, 234:1040–1057.

36. Lupas A, Dyke Mv, Stock J: Predicting Coiled Coils from Protein Sequences. *Science* 1991, 252:1162–1164.

37. Karlin S, Brendel V: Chance and Statistical Significance in Protein and DNA Sequence Analysis. *Science* 1992, 257:39–49.

38. Wootton JC, Federhen S: Statistics of Local Complexity in
• Amino Acid Sequences and Sequence Databases. *Comput Chem* 1993, 17:149–163.

An elaborate analysis of low-complexity regions in protein sequences and their detection is described. The approach extends beyond single sequences, and provides means for the generation of small databases.

39. Claverie J-M, States D: Information Enhancement Methods
• for Large Scale Sequence Analysis. *Comput Chem* 1993, 17:191–201.

The authors describe an algorithm and its implementation for detection and masking of compositionally biased regions in protein sequences.

40. Henikoff S, Henikoff JG: Automated Assembly of Protein Blocks for Database Searching. *Nucleic Acids Res* 1991, 19:6565–6572.

41. Henikoff S, Henikoff JG: Performance Evaluation of Amino
• Acid Substitution Matrices. *Proteins* 1993, 17:49–61.

Evidence is provided for the superior performance of the BLOSUM family of substitution matrices for database searching and sequence comparison.

42. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic Local Alignment Search Tool. *J Mol Biol* 1990, 215:403–410.

43. Gish W, States DJ: Identification of Protein Coding Re-
•• gions by Database Similarity Search. *Nature Genet* 1993, 3:266–272.

An extension of the BLAST suite of programs that allows the search of protein databases using a nucleotide as a query. Apart from its use for the identification of protein coding regions, this approach has other applications, such as the detection of frameshifts and the quality control of DNA sequencing.

44. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment. *Science* 1993, 262:208–214.

45. Rost B, Sander C: Prediction of Protein Secondary Structure at Better Than 70% Accuracy. *J Mol Biol* 1993, 232:584–599.

46. Piatigorsky J, Wistow GJ: Enzyme/Crystallins: Gene Sharing as an Evolutionary Strategy. *Cell* 1989, 57:197–199.

47. Bork P, Sander C, Valencia A: Convergent Evolution of Similar Enzymatic Function on Different Protein Folds: The Hexokinase, Ribokinase, and Galactokinase Families of Sugar Kinases. *Protein Sci* 1993, 2:31–40.

48. Bork P, Ouzounis C, Sander C, Scharf M, Schneider R, Sonnhammer E: What's in a Genome? *Nature* 1992, 358:287–287.

49. Koonin EV, Bork P, Sander C: Yeast Chromosome III: New
• Gene Functions. *EMBO J* 1994, 13:493–503.

The power of functional prediction by homology was demonstrated for yeast chromosome III. The various methods used set the percent-

age of functional assignments for yeast chromosome III proteins at 60%.

50. Bork P, Ouzounis C, Sander C, Scharf M, Schneider R, Sonnhammer E: Comprehensive Sequence Analysis of the 182 Predicted Open Reading Frames of Yeast Chromosome III. *Protein Sci* 1992, 1:1677–1690.

51. Prasad R, Widen SG, Singhal RK, Watkins J, Prakash L, Wilson SH: Yeast Open Reading Frame YCR14C Encodes a DNA β-Polymerase-Like Enzyme. *Nucleic Acids Res* 1993, 21:5301–5307.

52. Rohde K, Bork P: A Fast, Sensitive Pattern-Matching Approach for Protein Sequences. *Comput Appl Biosci* 1993, 9:183–189.

53. Crennel SJ, Garman EF, Laver WG, Vimr ER, Taylor GL: Crystal Structure of a Bacterial Sialidase (from *S. typhimurium* LT2) Shows the Same Fold as an Influenza Virus Neuraminidase. *Proc Natl Acad Sci USA* 1993, 90:9852–9856.

54. Glaser P, Kunst F, Arnaud M, Coudart M-P, Gonzales W, Hullo M-F, Ionescu M, Lubochinsky B, Marcelino L, Moszer I, *et al.*: ***Bacillus subtilis*** Genome Project: Cloning and Sequencing of the 97 Region from 325 to 333. *Mol Microbiol* 1993, 10:371–384.

55. Honoré N, Bergh S, Chanteau S, Doucet-Populaire F, Eiglmeier K, Garnier T, Georges C, Launois P, Limpaiboon T, Newton S *et al.*: Nucleotide Sequence of the First Cosmid from the ***Mycobacterium leprae*** Genome Project: Structure and Function of the Rif-Str Regions. *Mol Microbiol* 1993, 7:207–214.

56. Peterson SN, Hu P-C, Bott KF, Hutchison CAI: A Survey of the ***Mycoplasma genitalium*** Genome by Using Random Sequencing. *J Bacteriol* 1993, 175:7918–7930.

57. Waterston R, Martin C, Craxton M, Huynh C, Coulson A, Hillier L, Durbin R, Green P, Shownkeen R, Halloran N, *et al*: A Survey of Expressed Genes in ***Caenorbabditis elegans***. *Nature Genet* 1992, 1:114–123.

58. McCombie WR, Adams MD, Kelley JM, FitzGerald MG, Utterbeck TR, Khan M, Dubnick M, Kerlavage AR, Venter JC, Fields C: ***Caenorbabditis elegans*** Expressed Sequence Tags Identify Gene Families and Potential Disease Gene Homologues. *Nature Genet* 1992, 1:124–131.

59. Higgins DG, Bleasby AJ, Fuchs R: CLUSTAL V: Improved Software for Multiple Sequence Alignment. *Comput Appl Biosci* 1992, 8:189–191.

P Bork, C Ouzounis and C Sander, European Molecular Biology Laboratory, D-69012 Heidelberg, Germany.

Article No. mb982144 J. Mol. Biol. (1998) **283**, 707–725

JMB

REVIEW

Predicting Function: From Genes to Genomes and Back

Peer Bork*, Thomas Dandekar, Yolande Diaz-Lazcoz Frank Eisenhaber, Martijn Huynen and Yanping Yuan

European Molecular Biology Laboratory, Meyerhofstr. 1 PF 10.2209, D-69117 Heidelberg, Germany and Max-Delbrück-Centrum für Molekulare Medizin Robert-Rössle-Str. 10 D-13122 Berlin-Buch, Germany

**Corresponding author*

Predicting function from sequence using computational tools is a highly complicated procedure that is generally done for each gene individually. This review focuses on the added value that is provided by completely sequenced genomes in function prediction. Various levels of sequence annotation and function prediction are discussed, ranging from genomic sequence to that of complex cellular processes. Protein function is currently best described in the context of molecular interactions. In the near future it will be possible to predict protein function in the context of higher order processes such as the regulation of gene expression, metabolic pathways and signalling cascades. The analysis of such higher levels of function description uses, besides the information from completely sequenced genomes, also the additional information from proteomics and expression data. The final goal will be to elucidate the mapping between genotype and phenotype.

Keywords: genomes; computational tools; function prediction; comparative genome analysis; proteomics

Genomes and function prediction

Prediction of protein function using computational tools becomes more and more important as the gap between the increasing amount of sequences and the experimental characterization of the respective proteins widens (Bork & Koonin, 1998; Smith, 1998). With the availability of complete genomes we face a new quality in the prediction process (Table 1) as context information can be utilized when analysing particular sequences. This review focuses on the added value of genomic information on the many steps of function prediction from genomic sequence. The first reports on completely sequenced genomes give an excellent overview of the evolving state of the art in the analyses of particular genomes (Fleischmann *et al.*, 1995; Fraser *et al.*, 1995, 1998; Himmelreich *et al.*, 1996; Goffeau *et al.*, 1996; Kaneko *et al.*, 1996 Blattner *et al.*, 1997; Tomb *et al.*, 1997; Kunst *et al.*, 1997; Bult *et al.*, 1996; Smith *et al.*, 1997; Klenk *et al.*, 1997). In addition, there are numerous reviews that touch on the extraction of functional features from sequence (e.g. Bork *et al.*, 1994; Andrade *et al.*, 1997; Koonin & Galparin, 1997; Bork & Koonin, 1998), but very few reviews have been published that systematically summarize the additional information for function prediction that is provided by the presence of entirely sequenced genomes (original papers e.g. by Mushegian & Koonin, 1996a,b; Himmelreich *et al.*, 1997; Koonin *et al.*, 1997; Tatusov *et al.*, 1996, 1997; Huynen & Bork, 1998; Huynen *et al.*, 1997, 1998a; Dandekar *et al.*, 1998b).

What is function?

"Function" is a very loosely defined term that only makes sense in context. Most current efforts aim at predicting protein function, but there are other types of function, e.g. RNA function or organelle function, that also need to be explored. Even to describe "protein function" requires a broad range of attributes and features (Figure 1). Molecular features such as enzymatic activity, interaction

Present address: Y. Diaz-Lazcoz, Laboratoire Genome et Informatique; Batiment BUFFON, Universite de Versailles-Saint Quentin, 45, avenue des Etats-Unis, 78035 Versailles Cedex, France.

E-mail address of the corresponding author: Bork@EMBL-Heidelberg.de, Dandekar@EMBL-Heidelberg.de, Yolande.Diaz@genetique.uvsq.fr Eisenhaber@EMBL-Heidelberg.de, Huynen@EMBL-Heidelberg.de, Yuan@EMBL-Heidelberg.de

Table 1. Added features from complete genome analysis for function prediction

Genome specific patterns in the DNA and their usage in genome annotation	
Feature:	Genome-specific (poly)nucleotide frequencies, codon usage
Usage	→ Identification of genes → Identification of recent horizontal gene transfers into the genome
Feature:	Genome-specific signal sequences like regulatory regions, promotors
Usage	→ Gene identification, identification of the mode of regulation of genes, regulatory regions in mRNA, specification of the boundaries of genes → Operon identification
Usage of the complete set of genes in a genome and comparative genome analysis	
Feature:	The finding of orthologs by comparative genome analysis
Usage	→ Narrowing down the function of a gene → Identification of (conserved) regulatory signals neighbouring the orthologues
Feature:	Conserved genome organization
Usage	→ Genes in a conserved clusters have related functions, show physical interaction
Feature:	Differential genome analysis
Usage	→ Identification of the functions that are absent from a genome → If an orthologous gene is absent, but the function is present, missing genes point either to a wrong annotation or a non-orthologous gene transfer → Identification of the functions that are specific to a genome, and might be responsible for the species' specific phenotype, delineation of the mapping between genotype and phenotype → Correlation in the patterns of occurrence of genes in the comparison of multiple genomes points to functional relations between the genes
Feature:	Complete list of detected gene sequences
Usage	→ Identifying the optimal candidate gene in the whole genome for an observed enzymatic activity

Various types of patterns and (context) information that become available with the analysis of the complete genome can be used for function prediction at "lower levels", e.g. in the prediction of the function of single genes.

partners, and pathway context are currently being predicted, but only qualitatively. Expression patterns, regulation, kinetic properties, localization and concentration effects and, even more so, dysfunctions, environmental influence, fitness contribution or clinical symptoms can currently hardly be predicted. There is furthermore a relatively poor knowledge of the mechanisms of posttranslational modifications (Esko & Zhang, 1996). For example, although some sequence patterns for preferred glycolysation sites are known, the prediction accuracy is still limited and the assignment does not include the kind of sugar or carbohydrate that is attached, so that most of the functional features of the respective proteins will remain hidden.

The main goal will be to bridge the gap between genotype and phenotype (Figure 2), i.e. to understand the genotype to a degree that the phenotypic features can be predicted: What are the genes responsible for a certain disease phenotype and which proteins of the respective pathway (or an alternative one) are the best targets for a drug to be developed, or which variations at the DNA level are best suited for the respective diagnostics? Which genes have to be changed to achieve a desired phenotype? To answer such questions in a more general way, one needs a detailed understanding of the function of higher order processes, including the complex interaction between the heritable part of the phenotype and the environment. This will require a whole battery of novel types of experimental data with appropriate bioinformatics support.

Nevertheless, it is important to extract as much information as possible from sequence data using the already available (and inexpensive) computational tools to guide experimental work.

Functional prediction for gene products by annotation transfer from homologous sequences

When homologues of a query are identified in a database search (Bork & Gibson, 1996), the annotated information of the homologue and the taxonomic, biochemical and/or molecular-biological context of the query protein are used to extrapolate possible structural and functional features of the query protein. This approach has proven extremely successful although, from a formal point of view the hypotheses generated must be experimentally verified (Eisenhaber *et al.*, 1995). The information transfer from well-studied proteins to uncharacterized gene products has to be done carefully since (i) a similar sequence does not always imply similar protein structure (Sander & Schneider, 1991) or function (in particular in important details such as recognition loops) and (ii) the annotation of the database protein might be incomplete or even wrong.

Often (particularly in the case of automatic prediction programs), the function is transferred from another member in a multigene family, but not exactly from the functional counterpart in a different species. Even orthologues (see below) can differ functionally in various organisms. It should also be emphasized that generally only the molecular functions of a protein can be transferred by analogy (Figure 1); it is rather rare that a particular sequence motif strongly correlates with cellular functions as in the case of the DEATH-domain, which is mainly contained in apoptose signalling

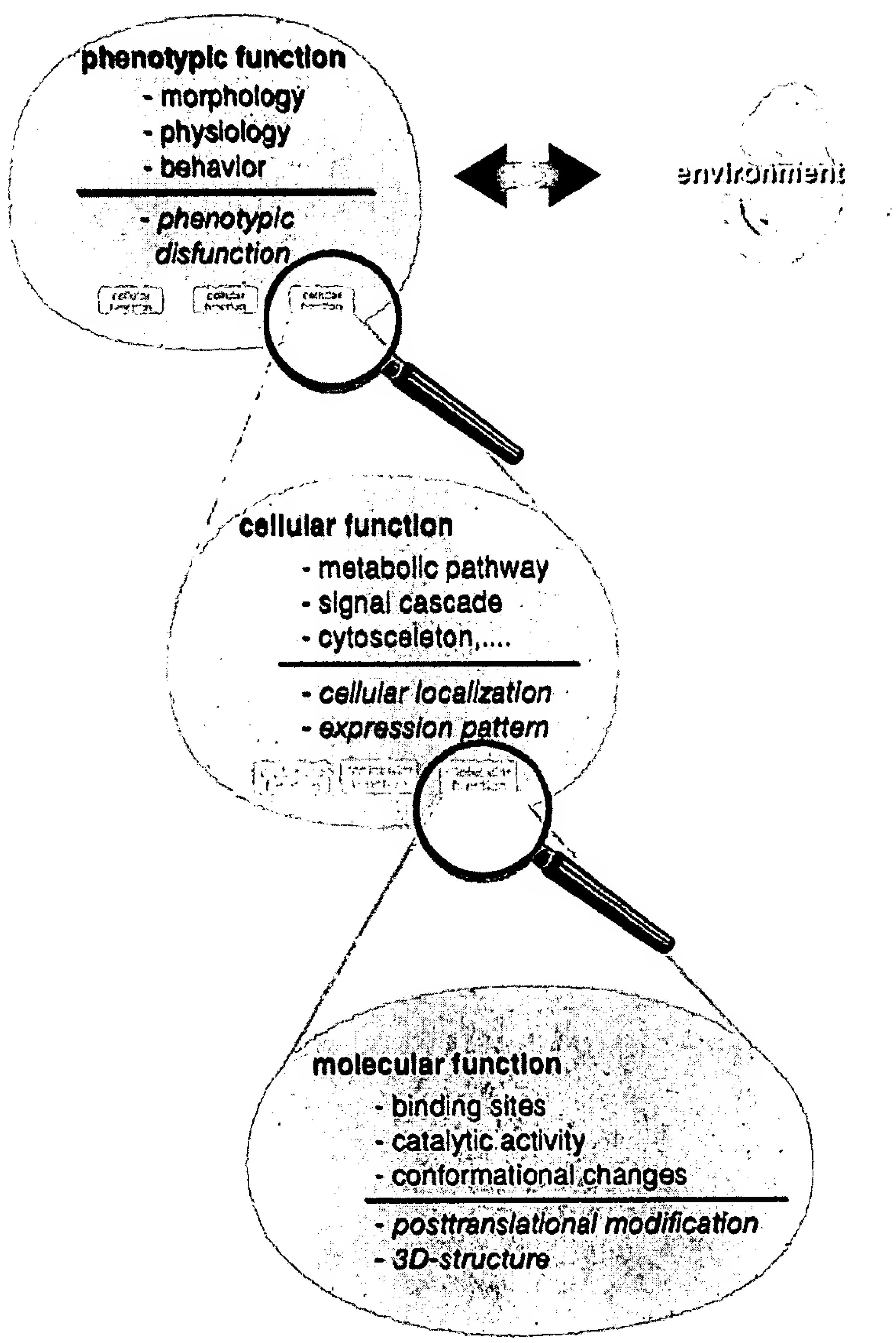


Figure 1. Characterization of protein function. Whereas nucleic acids fulfil the tasks of storage, transfer and processing of genetic information contained in the genome of living organisms, the proteins (gene products) form a complex (single- or multi-) cellular machinery for the realization of this genetic program (resulting in the phenotype) in dependency and in response to changing environment conditions. Therefore, protein function requires also a multilevel, hierarchical description comparable with the notions of primary, secondary, tertiary, and quaternary protein structure. Here, we propose a possible framework for functional characterization and, for each hierarchical level, both functional features and attributes are described. It should be noted that, for most proteins, a quantitative functional characterization is still a matter of the future and, today, a qualitative description of function for at least some hierarchical levels can be considered an achievement for many proteins. (1) Each protein has molecular (elementary) functions; e.g. it can have specific binding sites for substrates, low-molecular effectors, nucleic acids or other proteins. Given the set of allowed allosteric conformational changes, of possible interactions with other molecules, and of kinetic properties, etc., the protein can, for example, catalyse a metabolic reaction, it may transmit a signal to other proteins or DNA or be able to fit into cytoskeletal macromolecular associates. Structural properties of a protein are attributes for the execution of function, therefore, 3D structural information greatly facilitates understanding of function. Also, possible posttranslational modifications such as glycosylation, propeptide cleavage, or protein splicing are an important preposition for a protein to fulfil its molecular function. (2) A set of many co-operating proteins is responsible for a physiological (cellular) function (metabolic pathway, signal transduction cascade, structural associate etc.). The cellular function of a protein is always context-dependent and is characterized by taxon, organ, tissue, etc. Subcellular localization is an essential attribute for this level. For proper functioning, the protein has to be translocated to the correct intra- or extracellular compartments in a soluble form or to be attached to a membrane. All types of regulation of protein activity are another attribute. For example, the amount of protein molecules is often controlled *via* gene expression which might be limited to certain types of cells or tissues or to specific periods in the cell cycle or the individual ontogenese (expression pattern). (3) Finally, the totality of the physiological subsystems and their interplay with various environmental stimuli determines phenotype properties (phenotypic function), the morphology and physiology of the organism and its behaviour. Some phenotype properties may be traced to the activity of a single gene but most are determined by the co-operative action of many gene products. The absence of activity of a specific gene can result in phenotypic dysfunction. The knowledge of whole genomes will open a new era in the investigation of properties determined by many genes since the total set of genes influencing the phenotype is known.

proteins. Sometimes only the expression pattern and the tissue context determine the final functionality (for example, high sequence identity and even gene sharing between metabolic medium-chain dehydrogenases and eye lens crystallins; Piatigorsky & Wistow, 1991; Persson *et al.*, 1994; Serry *et al.*, 1998). Proteins (or more precisely, their domains) as structural and functional modules are multiply adapted by evolutionary processes and re-used in a different context. Thus, higher order

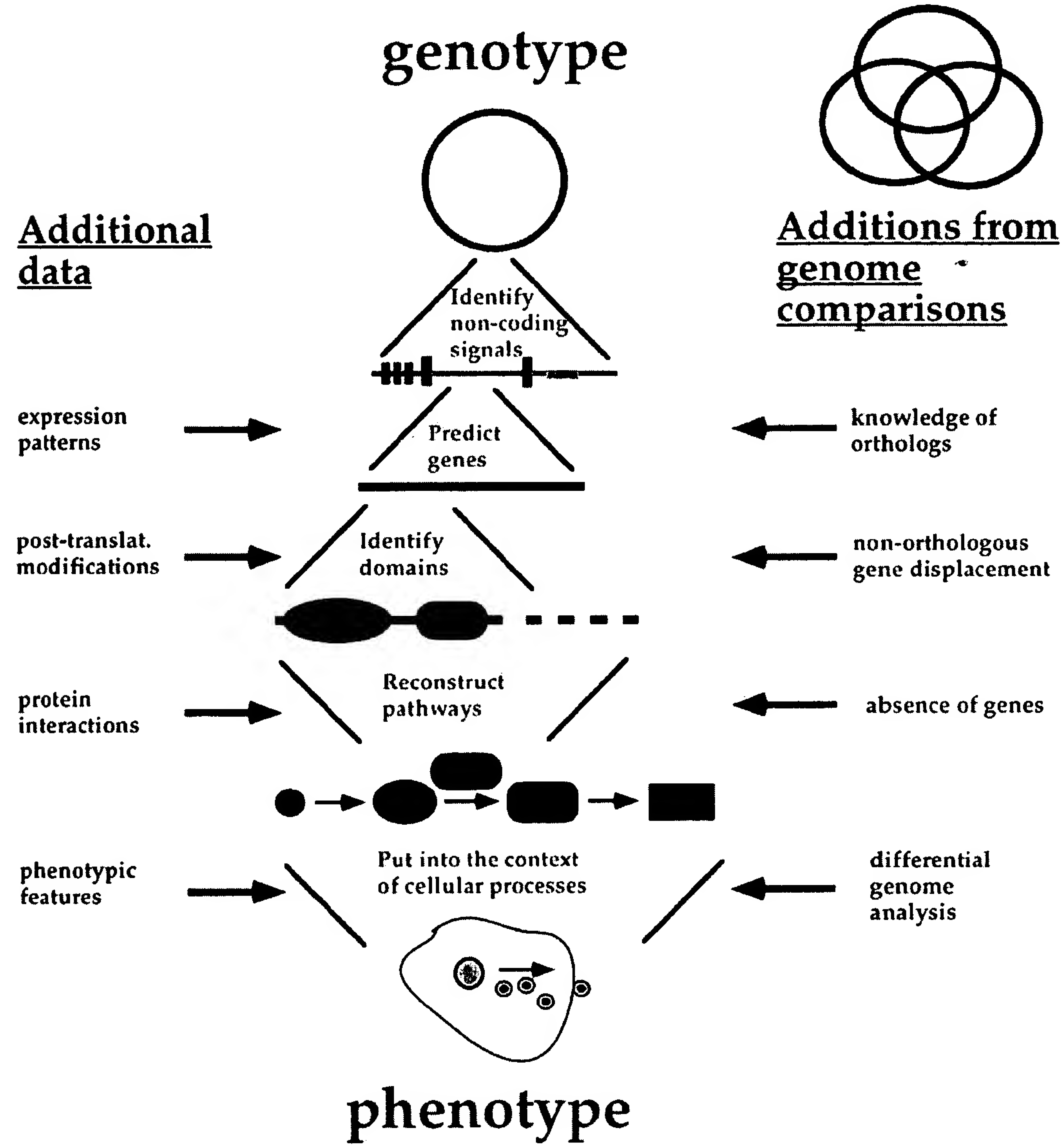


Figure 2. Function prediction scheme: zooming in and out. Whether gene, contig or genome: current methods concentrate on gene prediction and the annotation of individual genes that are then put into context. Due to our limited understanding of the genome, this is only possible by accessing complementary experimental information generated among others by proteomics research. Nevertheless, exploitation of genome information provides additional hints. This review follows the individual steps.

functions should be analysed in the biological context of the organism considered. Unfortunately, the functional knowledge of proteins reflected in their annotation (Figure 1) is frequently incomplete, sometimes erroneous or inconsistent, and often only cellular or even phenotypic functions are listed. For example, the human glia maturation factor (P17774) is described as growth factor (by definition extracellular!) but an in-depth sequence analysis revealed ADP-domains characteristic for cytoskeletal proteins (intracellular!).

Sequence and annotation quality in molecular databases

Function transfer by analogy requires knowledge about the quality of sequence data and functional annotation. Concerns have been raised about an accumulation (Bork & Bairoch, 1996) and even an explosion (Bhatia *et al.*, 1997) of errors in sequence databases.

In genome projects, two to tenfold sequence coverage is usually sampled. This is critical as

automated raw data acquisition (single read) is less than 99% accurate even when using optimized sequencers and software (Ewing *et al.*, 1998). Most of the ESTs (expressed sequence tags, i.e. single or sometimes double gel reads from cDNA) stored in current databases have much lower quality and require special caution as they are often also contaminated by cloning vectors or DNA of other sources including non-coding regions. More reasonable accuracy (at least over 99.85%) in all regions can be achieved only by systematic multiple coverage (Richterich, 1998). Nevertheless this will leave about one error per gene (mostly frameshifts) leading to considerable deviations at the protein level. Unless the accuracy is above 99.99% (the majority of the reading frames are sequenced without any error), a considerable error rate should be considered in the analysis.

Processing of raw genomic DNA includes identification of genes, their exon/intron structure, and the *"in silico"* translation into protein sequences by automatic methods. Given the limited accuracy of eukaryotic gene prediction methods (Burset & Guigo, 1996; Guigo, 1997, see below) and the impact of organelle- and species-specific translation tables, of pre- (RNA editing and splicing) and post- (propeptide cleavage and protein splicing, side-chain modifications) translational changes, the sequence quality of a given genomic segment is expected to be lower in protein databases than at the DNA level.

The value of sequences stored in databases is greatly increased by their functional annotation. However, automatic as well as manual annotations have all kinds of inaccuracies ranging from orthographic errors, simple spelling ambiguities, and incompleteness to semantic mistakes (Bork & Bairoch, 1996; Eisenhaber & Bork, 1998; Smith & Zhang, 1997). Function assignments obtained as a result of automatic homology searches are often not labelled as such and cannot easily be distinguished from true experimental data (Bork & Bairoch, 1996; Andrade & Sander, 1997). Furthermore, there is a gap between the current database annotation and the knowledge embodied in the scientific literature (Bork & Koonin, 1998).

Creating, updating, and correcting functional annotation is a costly effort absorbing a considerable amount of manpower. At the moment, there is no real alternative to manual input from experts. In the future, text analysis systems might support this process by automatically extracting abstracts of related articles from literature databases and selecting relevant keywords and text units for protein families (Guigo *et al.*, 1991; Guigo & Smith, 1993; Andrade & Valencia, 1997).

For analyses of genotype-phenotype relationships, the retrieval of complete sets of proteins from sequence databases with respect to their function is necessary. This can efficiently be achieved only by categorized protein function descriptions (Riley, 1998) for cellular (subcellular localization, involvement in metabolic pathways, signal transduction cascades, etc.) and phenotypic functions. However, functions are currently annotated in the form of plain text incorporating a large variety of vocabulary for the in-depth description of particular phenomena. Thus, they are not easily retrievable with keyword search engines such as SRS (Etzold *et al.*, 1996).

Computer-readable hierarchical systems of function description as envisioned in Figure 1 might be helpful, but controlled vocabularies such as in FLYBASE (ftp.ebi.ac.uk/pub/databases/edgp/misc/ashburner/fly_function_tree), the keywords in SWISS-PROT (expasy.hcuge.ch/sprot/), and, for catalytic functions, the system of Overbeek *et al.* (1997) put enormous pressure on the database curators. Such classifications also have to be adapted and updated frequently in accordance with the increasing understanding of the biological relationships.

Rule-based automatic algorithms that parse written annotations for defined questions might be a solution since a much smaller effort (compared with database reformatting) is required for their updates. For the deduction of cellular localization, a system of about 1000 biological rules was able to classify 88% of entries of SWISS-PROT (currently seen as one of the best annotated general protein sequence databases; Bairoch & Apweiler, 1998) into subcellular localization categories. This is considerable progress given that only 22% of the entries can be retrieved using querying stems of keywords such as "extracell" or "membrane" (Eisenhaber & Bork, 1998).

Annotating genomes

Function prediction usually starts with already assembled genomic or cDNA data: at best a complete genome (Figure 2). Several features intrinsic to DNA can be recognized first, before identification of genes and pathways, although detection of the latter enhances also the annotation of noncoding features in genomes.

Nucleotide frequencies

Nucleotide frequencies are one of the oldest features of genomes that have been studied, even before sequencing was available (Chargaff & Davidson, 1955). Biases in nucleotide frequencies exist both within and between genomes, they have various uses in gene and function prediction. In warm blooded vertebrates and angiosperms, for example, the genome is divided in regions, so called isochores, that differ in G+C content. Isochores with a high G+C content are relatively rich in genes (Saccone *et al.*, 1996). Biases in G+C content can hence be used to find genes. A number of bacterial species show biases in the nucleotide frequencies of the leading and lagging strands in replication (Mrazek & Karlin, 1998; Freeman *et al.*, 1998); these biases can be correlated with a bias in

the coding density, e.g. in *Bacillus subtilis* (Kunst *et al.*, 1997).

In the study of complete genomes, biases in nucleotide frequencies and codon usage provide an important clue for detecting recent horizontal transfers of genes into the genome (Figure 3; Medigue *et al.*, 1991). Variations in the codon usage can be described with a principle component analysis which divides the variation among orthogonal axes. Different axes correspond to independent sources of variation; the variation in codon usage that results specifically from horizontal gene transfer can be identified using auxiliary functional information from the genes (see below). In *Helicobacter pylori* for example, it is the first principle component that reflects horizontal gene transfer (Figure 3). On the basis of the variation in the codon usage in *E. coli* its genome has been predicted to consist of at least 10–15% of recently horizontally transferred genes (Medigue *et al.*, 1991). Biases of nucleotide frequencies within a genome also reveal information about their function apart from information about the evolutionary history of genes. Recently horizontally transferred genes are expected not to be involved in the core functions of the cell and to be relatively expendable (they were generally not present before the transfer). In *H. pylori* the regions with deviating nucleotide frequencies can be related to pathogenicity, or are prophages and/or are rich in insertion sequences (Figure 3). The same observation has been made in *Haemophilus influenzae* (Fleischmann *et al.*, 1995; Huynen *et al.*, 1997).

Repeats

For a large fraction of the DNA of multicellular eukaryotes no obvious function has yet been assigned. Most of it consists of repetitive elements. For example, *Alu* repeats may cover as much as 13% of the human genome (Mighell *et al.*, 1997). Repetitive, non-coding DNA should be filtered out as one of the first steps in function prediction to reduce the search space for the finding of genes in eukaryotic DNA (Jurka *et al.*, 1996). Coding regions contain repeats too, but these are hardly identifiable at the DNA level due to their divergence. They usually represent structural domains and should be detected at the protein level (see below). An exception are the trinucleotide repeats that are expanded in a number of disease genes (Chastian & Sinden, 1998); they can even specifically be used to search for such genes in DNA libraries (Pujana *et al.*, 1998).

In prokaryotes repeats are much less frequent. However, tetranucleotide repeats have been found in some virulence genes that increase variability by frameshift mutations (Hood *et al.*, 1996). More strikingly, repetitive elements even have been found in what are probably the smallest bacterial genomes, those of mycoplasmas. These have been hotspots for genome rearrangements *via* recombination, as can be deduced by whole genome comparison (Himmelreich *et al.*, 1997).

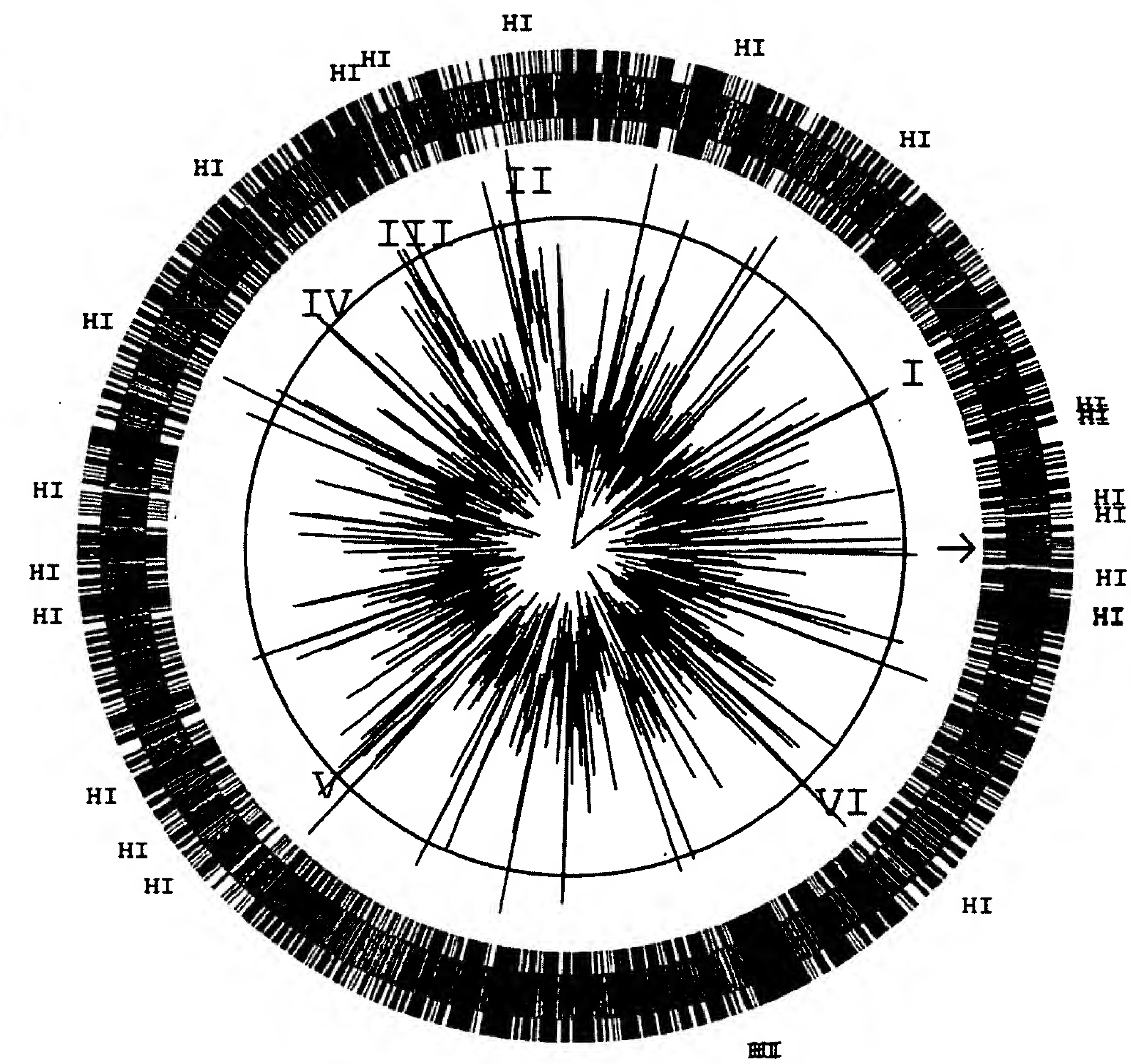
Regulatory regions

Regulatory regions can indicate when and how genes are expressed, repressed or co-expressed. Their computational detection is a powerful complement to novel experimental approaches (see proteomics, below). If known structures provide a template, simple consensus searches, matrix approaches and also programs taking into account specific features, structural constraints and energy values are available (reviewed by Dandekar & Scharma, 1998).

If no genomic template structures are available, neural networks (Demeler & Zhou, 1991; Pedersen & Engelbrecht, 1995; Ogura *et al.*, 1997), language based approaches (Trifonov, 1996) and other non-consensus search methods are important (e.g. Tiwari *et al.*, 1997). One can, for example, search for so-called CpG islands, which are, relative to the rest of the genome, abundant in the regulatory regions of mammalian housekeeping genes (Wirkner *et al.*, 1998). The combination of artificial *in vitro* evolution and genomic screening is another powerful way to identify a regulatory motif when no template structure or sequence is available. The computer based genomic screen delineates how close the *in vitro* selection procedure comes to the situation *in vivo* (Dandekar *et al.*, 1998a).

The challenge from complete genome sequences is double: first, a comprehensive annotation of known regulatory elements using specific searching methods (i.e. either templates for particular elements such as promotors, attenuators, terminators and enhancers or RNA secondary structure fitting methods; d'Aubenton-Carafa *et al.*, 1990; Brendel *et al.*, 1986); second, the identification of novel elements using comparative analysis and experimental indications (co-expression, etc.). Knowledge of gene expression and changes in gene expression patterns at a complete genomic level may revolutionize drug discovery processes. An overview of the complete genome allows much better tailoring of drugs and the discovery of correct, condition specific targets (Gelbert & Gregg, 1997).

A comparison of complete genomes identifies orthologous genes (see below). Their upstream regions can be screened for common regulatory signals in a much reduced search space. When co-expression patterns or functional interactions of genes are known, one can also search within the non-coding regions of a single genome. Unfortunately, regulatory regions in prokaryotes seem to be little conserved (Figure 4; Diaz-Lazcoz *et al.*, unpublished), thus it is necessary to include several species to increase the signal to noise ratio *via* multiple alignments. In the case of putative RNA structures, one can utilize methods that include base-pairing information (cf. Chan *et al.*, 1990; Han & Kim, 1993). These approaches, however, require



Helicobacter pylori

——	Genes with a homolog in the *E. coli* genome or a homolog in a non pathogenic *E. coli* plasmid
——	Genes (putatively) involved in interaction with the host
HI	Genes with a homolog in *H. influenzae* but not in *E.coli*
——	Others: e.g. metabolism, transposases, restriction enzymes
——	Genes with unknown functions
I - VI	clusters of genes with deviating codon usage
——	codon usage

Figure 3. Differential genome display of *H. pylori versus E. coli, H. influenzae*. The genes of *H. pylori* are divided into sets. Set 1 (green) are genes with a homologue in *E. coli*, set 2 (HI), genes with a homologue in *H. influenzae* but not in *E. coli*. Set 3 (red) are genes without a homologue in *E. coli* that are (putatively) involved in interaction with the host like virulence factors, outer membrane proteins and toxins. Set 4 (purple) are genes without a homologue in *E. coli* or *H. influenzae* that are not host interaction factors. A large fraction (63%) of the genes in *H. pylori* that have no homologue in *E. coli*, but for which some functional classification is possible, can be considered host interaction factors. The star-figure in the centre gives the values of the codon usage of the genes on the first principle component in distance to the centre. The first principle component corresponds roughly to the usage of A and to a lesser extent T in 3D codon positions. Hence, genes with a high value on this axis have a relatively high A+T content in third coding positions. Six clusters (I-VI) of at least three consecutive genes with, on average, the codon usage that deviates the most from the genomic mean were further analysed. The genes in these tend not to have any homologue in *E. coli* or *H. influenzae*, their closest relatives for which complete genome sequences are available. This observation supports that the genes in clusters I-VI result from horizontal gene transfer into the genome. Proteins from I and VI are hypothetical proteins with no known homologues other than proteins in *H. pylori* itself. Region II contains homologues of VirB4, a virulence factor and of transposases. Region III is the CAG pathogenicity island, whereas region V again is rich in transposases. Region IV consists of three proteins, HP0611-HP0613. Sequence analysis reveals a frameshift that would merge HP0611 with HP0612. The resulting protein is an ABC type 2 transporter, the only one that can be observed in *H. pylori*. ABC-2 transporters are involved in export of complex carbohydrates and play an important role in virulence.

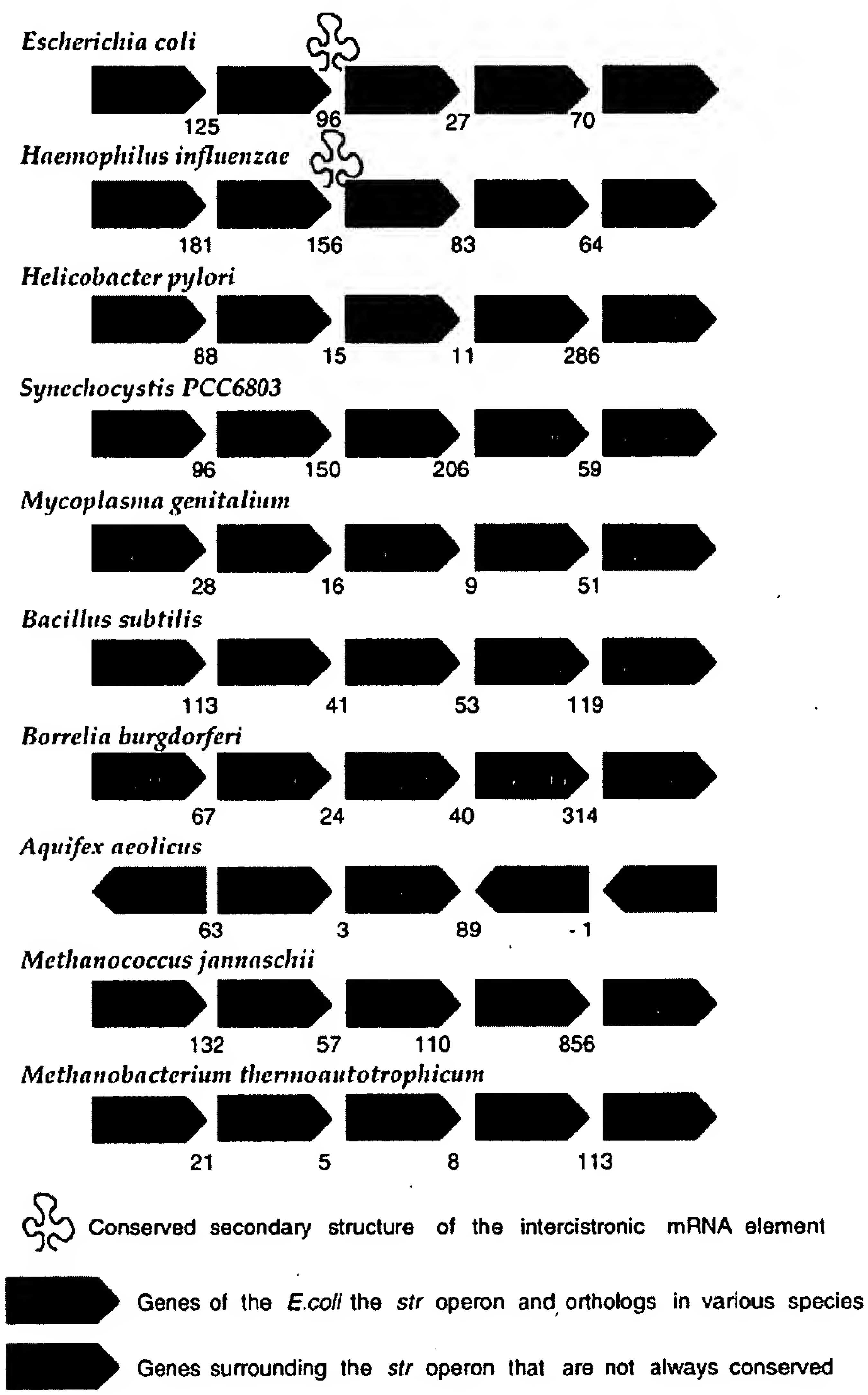


Figure 4. Variability of regulatory regions. Structure of the *str* operon and surrounding genes in eight Bacteria and two Archaea. The organisation of the *str* operon in *Archaeoglobus fulgidus* is essentially the same as in *M. jannaschii,* while the structure of the *str* operon in *Mycoplasma pneumoniae* is the same as that in *M. genitalium.* The arrows indicate the direction of transcription, the numbers under the arrows the lengths of the intergenic regions. Genes in green are orthologous to the genes in the same location in *E. coli.* Genes in light-red are shared at that location between two or more genomes other than *E. coli.* Gene names may vary in the official genome annotations, but were kept constant in this Figure for clarity purposes. Operon structure is generally not well conserved in prokaryotic evolution. The conservation of two genes besides each other in all the prokaryotic genomes that have been sequenced thus far can be regarded as an exception. Note that the conservation of the *str* operon does not follow the standard phylogenetic pattern: i.e. the operon structure in *E. coli* is more similar to that in the Archaea than it is to the operon structure in the closer related bacteria *B. burgdorferi* and *A. aeolicus.* In *E. coli* expression of the *str* operon is regulated by an RNA secondary structure located between the *rpsL* and *rpsG* genes (Saito & Nomura, 1994). A similar structure is present in *H. influenzae,* but is absent from the other species. Hence, regulatory elements appear even less conserved than gene order.

longer and relatively strong signals. Weaker motifs can be identified using statistical approaches without prior alignment of sequences (cf. Staden, 1989; Hertz *et al.*, 1990; Wolfertstetter *et al.*, 1996). Reliable statistics require, however, many orthologous sequences. The comparison based on orthologous regions in complete genomes, from different Gram-negative bacteria for instance, offers a new way to identify regulatory motifs without a preconception of the regulatory motifs revealed. In due course this and other approaches (see above) will improve quantitative predictions on expression and regulation in complete genomes and should also yield probabilities for tissue distribution of expression patterns and regulatory factors.

Gene prediction

The prediction of protein coding genes from DNA sequences can become a major bottleneck in genomics as currently there is quite a lot of information loss when genes cannot be identified correctly. In eukaryotes the situation is particularly complicated, due to the generally low coding density (probably as low as 2% in human) and the presence of introns surrounding the relatively short coding regions. Various different, but weak signals have to be combined such as promotors, splice sites, translational start and stop sites; different knowledge-based methods complemented by homology searches are applied to utilize them (Guigo,

1997). However, an analysis of the accuracy of all available packages for the prediction of coding sequences for a region of human DNA showed a low accuracy for the prediction of coding sequences and specifically the prediction of intron/exon boundaries (Burset & Guigo, 1996; Gelfand *et al.*, 1996; Guigo, 1997; Lukashin & Borodovsky, 1998).

In Archaea and Bacteria the situation is, relative to most eukaryotes, less complicated due to the almost complete absence of introns. In predicting protein coding regions, several methods make use of the information in the complete genome (Borodovsky *et al.*, 1994; Fraser *et al.*, 1998). Such bootstrapping methods use the genes that can easily be predicted, e.g. on the basis of the length of open reading frames and/or similarities to genes from other species to establish (i) taxon-specific patterns in codon usage, hexanucleotide frequencies and local complexity (information content), and (ii) taxon-specific signal sequences like poly(A) signals, regulatory sequences such as ribosome-binding segments (Shine-Delgarno segments) and promoters and start codons, etc. These patterns are currently implemented into Hidden Markov Models (HMMs) to predict the other genes in a genome. A method that relies on no *a priori* information to divide the genome into coding and non-coding regions has been shown to be successful (Audic & Claverie, 1998). These gene finding approaches should be complemented by another round of homology searches to find shorter or frameshifted genes that do not follow the codon usage of the organism.

Annotating individual proteins

Although homology searches are often already integrated into the gene prediction procedures, they are fully exploited only at the protein level with its higher sensitivity. Database searches are a standard technique for annotating proteins, but should be used in context with other methods (Bork & Koonin, 1998).

Domain analysis

Due to the modularity of many proteins, i.e. their multidomain architecture, the first step in functional annotation should be a scan for known domains in a query protein. Several databases exist that comprise patterns or profiles, i.e. fingerprints of already classified domains, and are well-suited for this first scan. Although somewhat redundant, they each have their individual strengths. PROSITE (Bairoch *et al.*, 1997) is one of the oldest and probably most widely used. It is well-annotated and covers more than 1000 different domains. With the inclusion of PROFILESCAN there is now also access to more than 250 domains that cannot easily be described with the classical PROSITE consensus string. A drawback, perhaps, is that the profiles are not yet fully integrated and most of them are not exhaustively annotated (which is a huge amount of work). BLOCKS (Henikoff *et al.*, 1998) is derived from PROSITE and offers ungapped alignments that are, in turn, used for a pattern matching approach which is more sensitive than the consensus string matching method of the original PROSITE database.

PRINTS describes a protein domain with a set of several motifs separated along the sequence (Attwood *et al.*, 1998). Version 17.0 is extensively annotated and comprises about 800 fingerprints with in total about 4500 motifs.

PFAM contains a collection of accessible multiple alignments that are translated into hidden Markov models; version 2.1 is a large collection covering 527 families that match at least once 47% of all SWISS-PROT entries in release 34 (Sonnhammer *et al.*, 1998); it more sensitive than the classical PROSITE or PRINTS, has poorer annotation, but has many entries crosslinked to other domain databases. SMART concentrates only on mobile domains and hence is not exhaustive, but has a high sensitivity and selectivity, takes care of domain borders and provides additional annotation features (Schultz *et al.*, 1998). Each of the databases offers search software on the web and there are efforts under way to overcome the difficulties of different formats and annotation styles.

Intrinsic feature analysis

Current database search techniques are all hampered by compositionally biased (low complexity) regions with a reduced residue alphabet. This includes (1) transmembrane regions: accumulations of ten hydrophobic residues in segments of length 20 of non-homologous transmembrane proteins are treated as homologous, (2) coiled coil segments (widespread heptarepeats with patterns of hydrophobic and polar residues) that pollute database search outputs with high scoring similarities to analogous (but probably not homologous) coiled coil regions in other proteins, (3) small repeats that lead to a bias in amino acid composition and (4) other regions with biases towards one or several amino acids such as proline-rich or glutamine-rich regions.

Methods exist for the identification of all those features. Most of these use many sequences with the feature as training sets and identify the feature knowledge based. A general method for finding low complexity regions, SEG (Wootton & Federhen, 1996) is already integrated into BLAST (Altschul *et al.*, 1997) in the form of a filter. Special types of composition bias can, of course, be predicted better by specialised methods such as coiled coil predictors (for a review see Lupas, 1997), transmembrane helix recognition (e.g. TOPPRED2, von Heijne, 1992) or even for subclasses of those such as signal sequences (e.g. SIGNALP, Nielsen *et al.*, 1997). For transmembrane regions, a variety of methods exists with widely varying outputs. It is worrying that when using different methods for

genome analysis the results vary greatly, e.g. the fraction of transmembrane proteins in *Mycoplasma genitalium* was predicted to be 18% (Fischer & Eisenberg, 1997), 24% (Koonin *et al.*, 1997), 30% (Arkin *et al.*, 1997) and 36% (Frishman & Mewes, 1997).

To avoid spurious hits and thus erroneous transfer of functional information, such regions can be filtered out, for example using the SEQ option in BLAST (Altschul *et al.*, 1997; replaced by "neutral" Xes for "any amino acid"). One has to bear in mind though that also such residues can contain useful functional and structural information and need to be annotated.

Another functional feature, especially in eukaryotic proteins, is the presence of posttranslational modifications. The EXPASY server provides useful software tools to detect and describe these (www.expasy.ch/www/tools.html).

Homology analysis

A classical database analysis should only be performed after identification and masking of domains and intrinsic features described above. This has the advantage of search space reduction and of better annotation quality. A database search using BLAST (Altschul *et al.*, 1997) or FASTA (Pearson, 1998) often reveals significant homologues, but this is then only the beginning of a complicated, and mostly manual transfer of functional information from the homologue in the database to the query sequence as one does not know how many of the functional features are shared (Doerks *et al.*, 1998). Averaged over all species, the chance that a newly sequenced gene has a homologue in sequence databases detectable by BLAST is already above 70% (e.g. 84% for yeast chromosome III; Bork & Koonin, 1998; 70–85% for Bacteria and 73% for Archaea; Koonin *et al.*, 1997, but lower for animals), while the fraction for which some functional features can be predicted is at least 70% in Archaea and Bacteria (Koonin *et al.*, 1997). For more than a third of all bacterial proteins, some homology-based fold assignments can be done with high confidence (Huynen *et al.*, 1998b; M.A.H. *et al.*, unpublished). Knowing the 3D structure of a protein is crucial in the understanding of the relation between sequence and function. In the case that amino acid identity levels to sequences with known 3D structures are higher than 50%, homology modelling can be used to further elucidate the roles and interactions of individual amino acids (Johnsson *et al.*, 1994; Eisenhaber *et al.*, 1995; Sanchez & Sali, 1997; Rodriguez & Vriend, 1997). Other predicted structural features such as secondary structure elements (Rost & O'Donoghue, 1997) can also be used in functional characterization. Characterization of a potential protein or RNA secondary structure can help to assess whether an open reading frame codes for a protein or a sequence codes for a functional RNA structure (Huynen *et al.*, 1996), respectively, or to test hypotheses based on other, independent observations.

Only in the minority of cases can functional and structural features of a homologue be transferred to the query sequence as is (see above, Figures 1, 2) because often only some of the features are shared. Functional equivalence is only likely for orthologues.

Finding orthologues

Orthologues (Figure 2, top right) are genes whose independent evolution reflects a speciation event rather than a gene duplication event (Fitch, 1970). They are likely to perform the same function in various species, and hence represent a refinement over homologues in sequence analysis and annotation. Knowledge of the complete genome and of its protein coding regions improves the detection of orthologues. Orthologues are expected to have the highest level of pairwise similarity between all the genes in two genomes (Tatusov *et al.*, 1996, 1997; Huynen & Bork, 1998), having diverged relatively recently compared to non-orthologous homologues. One needs to know all the proteins in two genomes to use relative levels of sequence identity to identify orthologues. Methods for the finding of orthologues rely both on relative similarity of genes from various genomes, and on information from the context of a gene in a genome. If two genes from different genomes share the same context, e.g. in the form of being a neighbour to a gene that also has the highest pairwise similarity between the two genomes, this supports them being orthologues of each other. The comparison of the sequence tree and the species tree can help in identifying orthologues (Yuan *et al.*, 1998), assuming that the genes have not been subject to horizontal transfer. Apart from information about the "functions" present in the genome, orthologues also provide information about the evolution of gene regulation. Specifically by comparing the 5′ and 3′ regions of orthologous genes one can obtain information about the evolution of promotors and operator/repressor sequences, and about the evolution of RNA secondary structures involved in gene regulation (see above). Orthologues should be the basis of subsequent reconstruction of pathways, rather than proteins for which we only know that they are homologous. Within the current databases, only a minor fraction of homologous relations can be classified as orthologous and thus one has to incorporate external data (Figure 2, left) for further function characterization.

Searching genes for a function

A tool that further exploits the information from comparing genomes for function prediction is differential genome analysis (Huynen *et al.*, 1997, 1998a). The genes that are not shared between two genomes are probably responsible for species-

specific phenotypes, as can be shown in the comparison of the pathogenic *H. influenzae* with the closely related but relatively benign *E. coli*. A large fraction (70%) of the genes in *H. influenzae* for which there are no homologues in *E. coli* and for which some functional annotation is possible can indeed be considered host interaction factors (Huynen *et al.*, 1997). Also in the pathogen *H. pylori* the fraction of genes that is not shared with *E. coli* is relatively enriched in host interaction factors (Figure 3). Taking differential genome analysis one step further one can show how gene content correlates with phenotype in multiple genome comparisons (Huynen *et al.*, 1998a). Although the correlations between gene content and phenotype cannot be used to predict the function of specific genes, they can serve as a filter to select genes that are probably responsible for specific functions. Or, in other words, to search for "genes for a function" rather than to search for "functions for a gene".

Incorporating proteomics data

Proteomics focuses on the protein products of the genome and their interactions rather than on DNA sequences (Humphery-Smith & Blackstock, 1997). It is thus complementary to the genomic and nucleic acid information (Kahn, 1995) exploiting novel tools such as 2D large scale analysis (Vietor & Huber, 1997) and powerful mass-spectrometry applications (Yates, 1998).

Protein identification and gene expression

Protein reading frames and expression behaviour in particular are not easy to predict from the genome sequence and profit from incorporation of additional experimental data (Figure 2, left). Co-expression as well as tissue- and organ-specific expression patterns at genomic scale are intensively studied (Hieter & Boguski, 1997; Zhang *et al.*, 1997) and recent techniques collect data on a genomic level.

Expressed sequence tag (EST) databases are available which contain information on gene expression that should correlate with the amount of redundancy, and on the tissue distribution of mRNA which can yield complex expression patterns (Boguski *et al.*, 1994; Zweiger & Scott, 1997). However, retrieval of this information is hampered by the high sequence error rate, by different splicing variants and by the often missing 5′ region necessary to determine the exact CDS start. Another caveat is that the EST approach has difficulties measuring genes with low expression.

Serial analysis of gene-expression (SAGE, Velculescu *et al.*, 1995) is a more rapid method to obtain partial sequence information from a very large set of expressed genes, e.g. differences in gene expression profiles in normal and cancer cells are identified by hundreds of differentially expressed transcripts, many of them growth factors (Zhang *et al.*, 1997). DNA chip-based gene-expression screening procedures are currently the fastest approach. Polymorphism with single base resolution is detected within minutes in the entire human mitochondrial genome (16.6-kilobases) by applying 135,000 probes simultaneously (array generated by light-directed chemical synthesis) and a two-colour fluorescent labelling scheme (Chee *et al.*, 1996). Systematic PCR of the entire yeast genome allows fluorescent readout of mRNA levels in different yeast environmental conditions such as changing glucose concentrations (DeRisi *et al.*, 1997; http://cmgm.stanford.edu/pbrown/explore). The correlation between mRNA and protein expression level is, however, debatable. Anderson & Seilhamer (1997) give a correlation coefficient 0.48 for expression levels in human liver measured either by two-dimensional electrophoresis (protein abundances) or by transcript image methodology (mRNA abundance measured by cDNA sequencing and cDNA clone count).

Direct determination of the major expressed proteins may thus be an independent and attractive alternative. The huge amount of work involved in this can today be substantially reduced by applying 2D gels and mass spectrometry and comparing experimental data to the annotated and predicted genome sequence. Link *et al.* (1997) identify the major part of the proteins and protein complexes from *H.influenzae* (300 out of 400 spots) after liquid chromatography (LC) and separation of the protein cleavage products of each 2D gel spot in a first mass spectrograph (MS) and further analysis in a second (LC/MS/MS approach). Several proteins not annotated in the genome sequence were identified by this approach.

Posttranslational modifications

After translation many proteins are further processed. This includes chemical modification of amino acids. Over 200 amino acid modification types are classified (Krishna & Wold, 1997), many more are expected (Annan & Carr, 1997). Such modifications are not apparent from the genome sequence, however, they are often critical for protein function. Two-dimensional gel electrophoresis coupled to mass spectrometry and modern software allows not only peptide mass fingerprinting for low quantities (Küster & Mann, 1998) but also specific detection of amino acid modifications on a large scale (Dongre *et al.*, 1997). For this, a database has to cover many of the reading frames likely to be encountered in the protein mixture analysed by the mass spectrometer. The EXPASY server (www.expasy.ch/www/tools.html) comprehensively links 2D gel experiments (e.g. separation from pH 4.0 to above 8.0 in the first dimension and from M_r 8-200 kDa in the second) to computer analysis tools. Nevertheless, determination of e.g. sugar modifications both by experiment and by software (e.g. EXPASY suite above) has limited accuracy, even including the kind of carbohydrate attached.

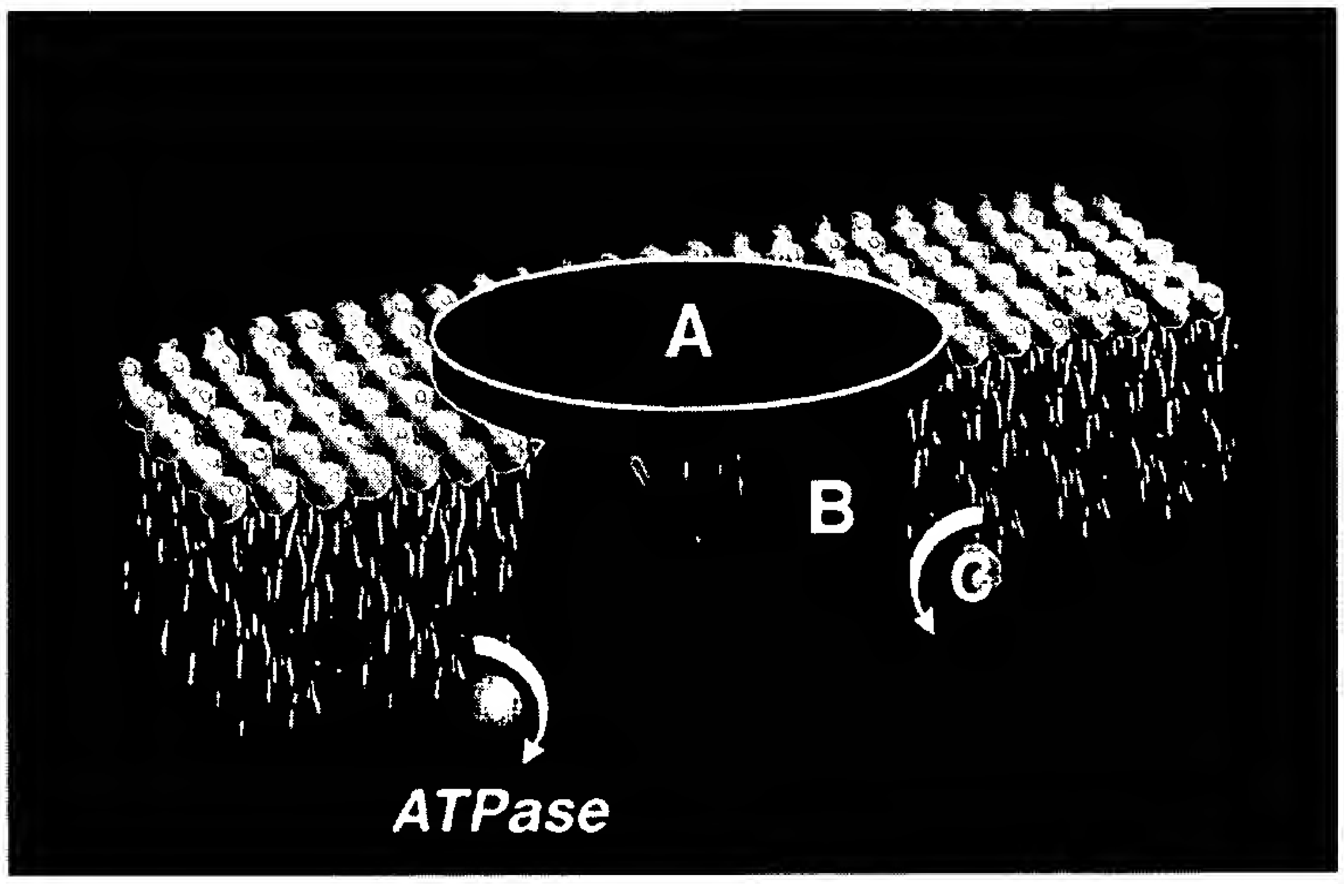


Figure 5. Protein interactions. Shown are ABC transporter proteins in the membrane of Gram-negative bacteria. Encoding genes are found as conserved gene clusters in the same sequential order in the complete genome sequences of *E. coli*, *H. influenzae* and *H. pylori*. They are an example of protein interactions predicted by triple comparison of complete genomes and additional confirmation by standard methods (see the text). A number of other protein interactions can also be suggested by comparative analysis of complete genomes.

Predicting function in higher order processes

Having predicted or determined functions for as many genes as possible and having assigned their interactions as well as their expression levels, it is a challenging task to put all the information into the context of cellular processes (Figure 1). A variety of databases and tools are emerging to support this procedure.

Information on tissue distribution

On the molecular level the processing machinery for metabolites differs in diverse tissues including absence of enzymes, receptors and structural proteins. On higher levels such as organ function, clinical impairment, drug metabolism or susceptibility to infections, tissue and phenotypic specific expression differences are key features of differentiation and help to find substances of therapeutical value. Data for humans are provided e.g. by TIGR (www.tigr.org/tdb/hgi/hgi.html), by NCBI (www.ncbi.nlm.nih.gov/UniGene/index.html and www.ncbi.nih.gov/dbEST/index.html), by SANBI-South African National Bioinformatics Institute (www.sanbi.ac.za/Dbases.html) and by the MRC human genetics unit (glengoyne.hgu.mrc.ac.uk). Such data should be used critically as low expression transcripts important for regulation such as tyrosine kinases may escape detection by EST sequencing or even Northern blots, and hence are misrepresented in databases. Techniques are still being improved (e.g. DNA chips (Brown, 1994)) and many data are not yet on the Web or are even completely inaccessible (e.g. in companies).

Analysis of protein interactions

Prediction and analysis of protein interaction uses both experimental (e.g. antibody precipitation, Maniatis *et al.*, 1989) and theoretical approaches. Two hybrid screening systems (Tsukamoto *et al.*, 1997) allow large scale screening, e.g. for 20 residue peptide sequences that correctly recognize (so called "aptamers") and inhibit cyclin-dependent kinase 2 (Colas *et al.*, 1997). Automation (with a considerable error rate though) and matching the data gathered with context and information such as common pathways is possible (Brent & Finley, 1997). Logical connections of protein interactions (e.g. with ras protein) can be revealed by a careful choice of reporter plasmids (Xu *et al.*, 1997).

A new way to identify protein interactions, comparative analysis between genomes, has revealed that the conservation of gene order between genomes with less than 50% protein identity is limited to those genes that code for proteins that physically interact with each other (Dandekar *et al.*, 1998b). Protein candidates for physical interaction that are identified by the conservation of their gene order can further be analysed by the methods mentioned above. An example are the ABC transporters which were experimentally shown to consist of physically interacting proteins (Eym *et al.*, 1996) and are found in conserved gene clusters in different genomes (Figure 5). The conservation of gene order can of course be used for the prediction of functional features of hypothetical proteins (interaction with a neighbour and, if this one is characterized, even participation in a pathway).

Reconstruction of pathways

The prediction of reactions and pathways (example: Figure 6) of the respective organisms integrates all the data above (including errors at different levels!) into its phenotypic context and yields a more complete picture of the biochemical and adaptive capabilities of the sequenced organism (Overbeek *et al.*, 1996). Mispredictions, wrong annotations and higher level errors (substrate specificity etc.) have to be minimized by context information and additional experimental data. Problems specific to pathway predictions arise, such as non-orthologous displacements (enzymatic

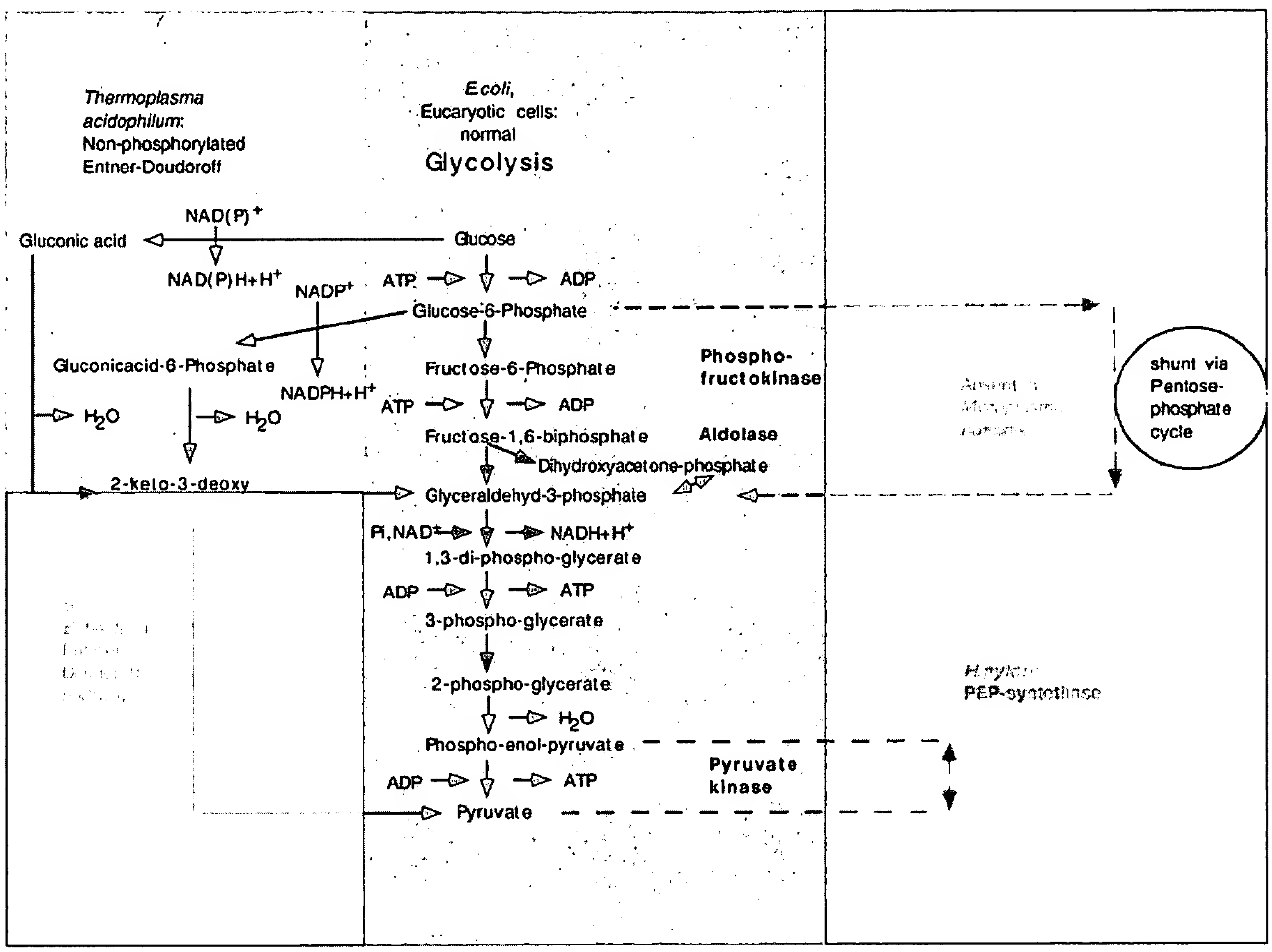


Figure 6. Prediction of metabolic pathways and pathway alignment. The glycolytic pathway (centre) and alternative other routes (sides) predicted from the genome and observed in several microorganisms are shown and compared (pathway alignment) to illustrate the often underestimated variability of metabolic pathways. Key enzymes discussed are shown in bold. In the glycolytic pathway (centre) two molecules of triose are derived from one hexose (as dihydroxyacetone phosphate can also be converted into glyceraldehyde-3-phosphate), the energy yield is two mole ATP per mole glucose. Genome analysis shows that the complete glycolytic pathway is present in *E. coli*, as it is, incidentally, in most eukaryotic organisms and cells including all human cells. In contrast, in *H. pylori* , a causative agent for stomach ulcer and chronic ulcerative gastritis, phosphofructokinase in the upper part of the glycolytic pathway and the important enzyme pyruvate kinase in the lower part seem to be missing. Thus a different route has to be taken in *H. pylori*. According to our analysis (right bottom), a homologue of phosphoenol-pyruvate (PEP) synthetase is present, which may support the missing step of the pyruvate kinase albeit at a reduced energy yield. The taking over of the role of pyruvate kinase by phophoenol-pyruvate (PEP) synthetase could be an adaptation to the highly acidic environment of the stomach in which *H. pylori* has to survive. More, and also more complex phenotypic features of *H. pylori* can be understood in this way by a pathway analysis utilizing differential genome comparisons (Huynen *et al.*, 1998a). What about alternatives for the first part of glycolysis? This is illustrated in Figure 6 for *Mycoplasma* species which are non-glycolytics: Phosphofructokinase is missing (as probably is the case in *H. pylori* where only some homologue to pfkB from *E. coli* is present but is likely to be utilized differently) and also aldolase is absent, for instance in *Mycoplasma hominis*. These species seem to channel instead glucose by the pentosephosphate cycle (Pollack *et al.*, 1997), which also yields glyceraldehyde-3-phosphate plus ribose and NADPH for nucleotide synthesis and is thus less dispensable than parts of glycolysis in these very compact genomes (Himmelreich *et al.*, 1997). Our own investigations indicate that this should only be stochiometrically possible if there are additional enzymes in the genome or additional functions of known enzymes which serve to replenish the pool of sugar phosphates in the pentose phosphate pathway. There are several further alternatives for converting glucose to pyruvate. The Entner-Doudoroff pathway (bottom left), is used instead of glycolysis in some bacteria (Danson & Hough, 1992). Furthermore, genome analysis by us and others shows that this route is present as a backup pathway for instance in all Gram-negative genomes analysed to date. The ATP yield is only one mole per mole glucose. Probably it survived as an exclusive pathway in some genomes due to its simplicity and direct yield of NADPH. Top left shows the non-phosphorylated Entner-Doudoroff-pathway. This is an example of paleo-metabolism and due to the direct conversion of glucose to gluconic acid not yet optimized to obtain any net ATP yield per mole glucose (Melendez-Hevia *et al.*, 1997). It is present in some Archaea such as *Thermoplasma acidophilum* (Fields, 1987).

activities may then be overlooked in homology searches; Koonin *et al.*, 1996).

Databases increasingly facilitate the prediction of metabolic pathways, notably EcoCyc (Encyclopedia of *E. coli* Genes and Metabolism), HinCyc: (Encyclopedia of *H. influenzae* Genes and Metabolism), PUMA (see below), Biocatalysis/Biodegradation Databases, Enzyme Database (e.g. www.expasy.ch/sprot/enzyme.html), Ligand Databases (e.g. at the Japanese genome net, its compound section is a collection of metabolic compounds including substrates, products, and inhibitors; www.genome.ad.jp/dbget/ligand.html), Klotho, the biochemical compounds declarative database; KEGG (Kyoto Encyclopedia of Genes and Genomes page) and pathway pages on the Web such as the Boehringer Mannheim pathways chart, NetBiochem Welcome Page or pages on particular organisms such as for soybean metabolism (cgsc.biology.yale.edu/metab. html). Experimentally verified pathway databases have been collected for regulatory circuits such as cell cycle in yeast, human and budding yeast (BRITE project, www.genomes.ad.jp/brite/Cellcyclemaps.html), as cross-references (object oriented database management system ACEDB) between protein kinases, their interactions, 3D structure and pathways by Igarashi & Kaminuma (1997) and for fly genes involved in pattern formation (Jacq *et al.*, 1997).

Several software tools reconstruct metabolic pathways, usually in association with databases (see above). Early efforts (Seressiotis & Bailey, 1988; Mavrovouniotis *et al.*, 1990) required extensive pre-analysis of the genome and the proteins encoded therein. More recent developments include Magpie (Multipurpose Automated Genome Project Investigation Environment), an automated genome analysis tool (Gaasterland & Sensen, 1996) that accesses several databases through an object and attribute viewer. Reaction equations, and compounds are taken from the Enzyme and Metabolic Pathway Database (Selkov *et al.*, 1996) and have been assigned *via* homology to proteins from several organisms. The precomputed reconstruction can be accessed *via* the Web. The WIT (What Is There) system (Overbeek *et al.*, 1997) is similar in concept, but offers a wide range of query options. It is a useful toolkit (http://www.cme.msu.edu/WIT) to briefly check for pathways that might be present in the genome of interest. Also with the KEGG database pathway computations are possible, for instance testing the completeness of an enzyme list (e.g. from a genome sequencing project) with regard to a certain pathway (Ogata *et al.*, 1998).

Nevertheless, current reconstruction of metabolic pathways from sequence is mostly done manually using various tools that guide the decisions with consideration of accumulated biochemical and biological knowledge. For example the EcoCyc WWW Server (Karp *et al.*, 1998) is used as a reference and each possible hit there is carefully checked for orthology (the whole protein function should be similar, the sequence similarity should not be restricted only to a functional domain; otherwise no complete function transfer possible). For the latter an efficient tool is the COGs server (clusters of orthologous genes; Tatusov *et al.*, 1997; http://www.ncbi.nlm.nih.gov/COG/). Profile alignments of important enzymatic activities such as signatures for pathways are also used and are being developed in several other laboratories (e.g. Rawlings & Searls, 1997).

The prediction of interdependencies of genes and metabolism

Utilizing the tools and approaches above together with methods for comparative sequence and genome analysis, a number of specific predictions in recently sequenced prokaryotic organisms have been made that go beyond the analysis provided in the publications on sequenced genomes for prokaryotes (Selkov *et al.*, 1996, 1997; Tatusov *et al.*, 1996; Koonin *et al.*, 1996; Strauss & Falkow, 1997) and eukaryotes (Oliver, 1997; Palsson, 1997). The plasticity and the enzyme variety even of very basic pathways turns out to be surprisingly high. Figure 6 illustrates this for variants from standard glycolysis encountered after genome analysis.

Predictions for protein functions and enzyme pathways just cover the repertoire of functions present. However, metabolic control analysis also considers quantitative aspects such as flux, flow, concentrations, stochiometric and allosteric effects, compartimentalization and regulation (see e.g. Schuster, 1996; Thomas & Fell, 1996; Bish & Mavrovouniotis, 1998 and references therein). Knowledge of possible metabolite flows (i.e. different paths and orders of reactions given a constant number of enzymes; "elementary modes", Schuster & Hilgetag, 1994, 1995; Liao *et al.*, 1996; Nuño *et al.*, 1997; Bonarius *et al.*, 1997) should improve the understanding of the context of identified enzymes in the near future. This requires well-studied systems. However, exactly these can be achieved by extensive genome and proteome analysis.

Comparative analysis of complete genomes provides further tools to study gene interdependence. For example, genes that depend on each other are expected to occur together in genomes or to be absent altogether. By doing large scale comparative genome analysis such correlations between genes become apparent and provide an extra tool for finding connections in metabolism or signalling cascades. An example are sets of genes shared by *M. genitalium* and one of either *M. jannaschii* (set 1) or *M. thermoautotrophicum* (set 2), but not by the other. Set 1 encodes among others, the functionally related proteins phosphoglucose isomerase, glyceraldehyde 3-phosphate dehydrogenase and pyruvate kinase, that are all involved in glycolysis, whereas set 2 contains the genes for DnaK and DnaJ, parts of a chaperone pathway (Huynen & Bork, 1998).

Robustness, modularity and interdependence

When considering all the levels discussed there seems to be a discrepancy between the complex nature of the networks of genes and their interdependence (e.g. *via* regulation) on the one hand and the surprising robustness (e.g. horizontal gene transfer or gene loss) on the other. One way in which such robustness might be achieved is a highly modular organisation, the interdependencies of genes would then be limited to small sets. As yet we do not have a quantitative understanding of the modularity of cellular organisation, including the genome, and its implications for the flexibility and robustness of evolution. One also needs to keep in mind that the examples of robustness we see are a selected set: evolution does not report negative results. We have tried to show here the powers of using information contained in entire genomes, i.e. context information and the interdependencies of genes within a genome. These rules affect the function prediction process in various ways.

Limited prediction accuracy at all levels and interdependence

Although many methods exist for various aspects of each prediction step, one has to bear in mind that they are not perfect and have only a limited accuracy. In addition, most of the methods have (sometimes hidden) parameters that influence the search result drastically (just switch in BLAST the matrix from default BLOSUM62 to PAM250 and watch the changes in the output). Fortunately, the loss of information in each step is compensated by the fact that data are produced by experimental methods in all the different levels (Figure 2). Thus, the errors do not add up and can be compensated by information from different levels e.g. by using genome information to improve the prediction of protein function as described here. Experimental validation of hypotheses can also be conducted at all levels, the interdependence allows even the interpretation of cellular data for molecular features and *vice versa*.

Modularity at each level and robustness

Modularity already is present at the DNA sequence level in repeats, ubiquitous promotors, duplicated segments, etc. The limited set of domains, used again and again as structural and functional scaffold, documents modularity at the gene and protein level. Displacement of non-homologous but functionally equivalent enzymes and the distinct pathway variants that all lead to the same compounds (see Figure 6) are evidence for modularity at the cellular level. Complex systems such as the cytoskeleton (animals *versus Mycoplasmas*) or even specialized organs (vertebrate *versus* octopus eye) do not represent unique solutions and reveal that even tissues can be re-invented on the basis of lower level modules. Thus, a remarkable robustness can be observed at all levels, the balance of which might seem surprising given the shuffling, horizontal transfer, disruption, insertion etc. of genetic material. On the other hand, the robustness represents also hope that functional features are more significantly implicated and predictable from sequence than previously expected.

Prediction of function from sequence is a considerably more complex enterprise than a simple sequence database search which represented the entire repertoire of tools a few years ago. In particular, with the arrival of multiple entirely sequenced genomes and experimental input at various complexity levels we have the chance to approach a new quality of understanding of cellular processes and their evolution.

Acknowledgements

The order of the authors is alphabetically. This work was supported by Deutsche Forschungsgemeinschaft (Bo 1099/3-1) and BMBF (grants 01KW9602/6; 0311748; 0311617). We thank Enrique Morrett and Shamil Sunyaev for critical reading of the manuscript and David Thomas for stylistic corrections. Most of all we acknowledge the work and efforts of all our colleagues who could not be mentioned in this review due to limitations of space and time and the limited selection such a review necessarily has to make.

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped Blast and PSI-Blast, a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.

Anderson, L. & Seilhamer, J. (1997). A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis*, **18**, 533–537.

Andrade, M. A. & Sander, C. (1997). Bioinformatics: from genome data to biological knowledge. *Curr. Opin. Biotechnol.* **8**, 675–683.

Andrade, M. & Valencia, A. (1997). Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts. *Intelligent Systems Mol. Biol.* **5**, 25–32.

Andrade, M., Casari, G., de Daruvar, A., Sander, C., Schneider, R., Tamames, J., Valencia, A. & Ouzounis, C. (1997). Sequence analysis of the *Methanococcus jannaschii* genome and the prediction of protein function. *Comput. Appl. Biosci.* **13**, 481–483.

Annan, R. S. & Carr, S. A. (1997). The essential role of mass spectroscopy in characterization of protein structure: mapping post-translational modifications. *J. Protein Chem.* **16**, 391–402.

Arkin, I. T., Brunger, A. T. & Engelman, D. M. (1997). Are there dominant membrane protein families with a given number of helices?. *Proteins: Funct. Genet.* **28**, 465–466.

Attwood, T. K., Beck, M. E., Flower, D. R., Scordis, P. & Selley, J. N. (1998). The PRINTS protein fingerprint database in its fifth year. *Nucl. Acids Res.* **26**, 304–308.

d'Aubenton-Carafa, Y., Brody, E. & Thermes, C. (1990). Prediction of rho-independent *Escherichia coli* transcription terminators. A statistical analysis of their RNA stem-loop, structures. *J. Mol. Biol.* **216**, 835–858.
Audic, S. & Claverie, J.-M. (1998). Self-identification of protein-coding regions in microbial genomes. *Proc. Natl Acad. Sci. USA*, **95**, 10026–10031.
Bairoch, A. & Apweiler, R. (1998). The SWISS-PROT protein sequence databank and its supplement TrEMBL. *Nucl. Acids Res.* **26**, 38–42.
Bairoch, A., Bucher, P. & Hofmann, K. (1997). The PROSITE database, its status and progress. *Nucl. Acids Res.* **25**, 217–221.
Bhatia, U., Robison, K. & Gilbert, W. (1997). Dealing with database explosion: a cautionary note. *Science*, **276**, 1724–1725.
Bish, D. R. & Mavrovouniotis, M. L. (1998). Enzymatic reaction rate limits with constraints on equilibrium constants and experimental parameters. *Biosystems*, **47**, 37–60.
Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., *et al.* (1997). The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1462.
Boguski, M. S., Tolstoshev, C. M. & Bassett, D. E., Jr (1994). Gene discovery in dbEST. *Science*, **265**, 1993–1994.
Bonarius, H. P. J., Schmid, G. & Tramper, J. (1997). Flux analysis of underdetermined metabolic networks: the quest for the missing constraints. *Trends Biotechnol.* **15**, 308–314.
Bork, P. & Bairoch, A. (1996). Go hunting in sequence databases but watch out for the traps. *Trends Genet.* **12**, 425–427.
Bork, P. & Gibson, T. J. (1996). Applying motif and profile searches. *Methods Enzymol.* **266**, 162–184.
Bork, P. & Koonin, E. V. (1998). Predicting function from protein sequence: Where are the bottlenecks? *Nature Genet.* **13**, 313–318.
Bork, P., Ouzounis, C. & Sander, C. (1994). From genome sequences to protein function. *Curr. Opin. Struct. Biol.* **4**, 393–403.
Borodovsky, M., Koonin, E. V. & Rudd, K. E. (1994). New genes in old sequence: a strategy for finding genes in the bacterial genome. *Trends Biochem. Sci.* **19**, 309–313.
Brown, P. O. (1994). Genome scanning methods. *Curr. Opin. Genet. Dev.* **4**, 366–373.
Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J. F., Adams, M. D., Reich, C. I., *et al.* (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science*, **273**, 1058–1073.
Brendel, V., Hamm, G. H. & Trifonov, E. N. (1986). Terminators of transcription with RNA polymerase from *Escherichia coli*: what they look like and how to find them. *J. Biomol. Struct. Dynan.* **3**, 705–723.
Brent, R. & Finley, R. L., Jr (1997). Understanding gene and allele function with two-hybrid methods. *Annu. Rev. Genet.* **31**, 663–704.
Burset, M. & Guigo, R. (1996). Evaluation of gene structure prediction programs. *Genomics*, **34**, 353–367.
Chan, L., Zuker, M. & Jacobson, A. B. (1990). A computer method for finding common base paired helices in aligned sequences: application to the analysis of random sequences. *Nucl. Acids Res.* **19**, 353–358.
Chargaff, E. & Davidson, N. J. (1955). *The Nucleic Acids*, Academic Press, New York.
Chastian, P. D. & Sinden, R. R. (1998). CTG repeats associated with human genetic disease are inherently flexible. *J. Mol. Biol.* **275**, 405–411.
Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X. C., Stern, D., Winkler, J., Lockhart, D. J., Morris, M. S. & Fodor, S. P. (1996). Accessing genetic information with high-density DNA arrays. *Science*, **274**, 610–614.
Colas, P., Cohen, B., Jessen, T., Grishina, I., McCoy, J. & Brent, R. (1997). Genetic selection of peptide aptamers that recognize and inhibit cyclin-dependent kinase 2. *Nature*, **380**, 548–550.
Dandekar, T. & Sharma, K. (1998). *Regulatory RNA*, Springer Verlag, Heidelberg.
Dandekar, T., Beyer, K., Bork, P., Kenealy, M.-R., Pantopoulos, K., Hentze, M. W., Sonntag-Buck, V., Flouriot, G., Gannon, F., Keller, W. & Schreiber, S. (1998a). Systematic genomic screening and analysis of mRNA in untranslated regions and mRNA precursors: combining experiment and computational approaches. *Bioinformatics*, **14**, 271–278.
Dandekar, T., Snel, B., Huynen, M. A. & Bork, P. (1998b). Conservation of gene order: a fingerprint of physically interacting proteins. *Trends Biochnol. Sci.* **23**, 324–328.
Danson, M. J. & Hough, D. W. (1992). The enzymology of archaebacterial pathways of central metabolism. *Biochem. Soc. Symp.* **58**, 7–21.
Demeler, B. & Zhou, G. (1991). Neural network optimization for *E. coli* promoter prediction. *Nucl. Acids Res.* **19**, 1593–1599.
DeRisi, J. L., Vishwanath, R. I. & Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.
Doerks, T., Bairoch, A. & Bork, P. (1998). Protein annotation: detective work for function prediction. *Trends Genet.* **14**, 248–250.
Dongre, A. R., Eng, J. K. & Yates, J. R. (1997). Emerging tandem-mass-sepctroscopy techniques for the rapid identification of proteins. *Trends Biotechnol.* **15**, 418–425.
Eisenhaber, F. & Bork, P. (1998). Wanted: subcellular localization of proteins based on sequence. *Trends Cell Biol.* **8**, 169–170.
Eisenhaber, F., Persson, B. & Argos, P. (1995). Protein structure prediction: recognition of primary, secondary, and tertiary structural features from amino acid sequence. *CRC Crit. Rev. Biochem. Mol. Biol.* **30**, 1–94.
Esko, J. D. & Zhang, I. (1996). Influence of core protein sequence on glycosamino-glycan assembly. *Curr. Opin. Struc. Biol.* **6**, 663–670.
Etzold, T., Ulyanov, A. & Argos, P. (1996). SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.* **266**, 114–128.
Ewing, B., Hiller, L., Wendl, M. C. & Green, P. (1998). Base-calling of automated sequncer traces using phred. I. Accuracy assesment. *Genome Res.* **8**, 175–185.
Eym, Y., Park, Y. & Park, C. (1996). Genetically probing the regions of ribose-binding protein involved in permease interaction. *Mol. Microbiol.* **21**, 695–702.
Fields, J. H. A. (1987). Fermentative adaptations to the lack of oxygen. *Can. J. Zool.* **66**, 1036–1040.
Fischer, D. & Eisenberg, D. (1997). Assigning folds to the proteins encoded by the genome of *Mycoplasma*

genitalium. *Proc. Natl Acad. Sci. USA*, **94**, 11929–11934.

Fitch, W. (1970). Distinguishing homologous from analogous proteins. *Syst. Zool.* **19**, 99–113.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J. D., *et al.* (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Nature*, **269**, 496–512.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, J. L., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., *et al.* (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science*, **270**, 397–403.

Fraser, C. M., Casjens, S., Huang, W. M., Sutton, G. G., Clayton, R., Lathigra, R., White, O., Ketchum, K. A., Dodson, R., Hickey, E. K., Gwinn, M., Dougherty, B., Tomb, J. F., Fleischmann, R. D., Richardson, D., *et al.* (1998). Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature*, **390**, 580–586.

Freeman, J. M., Plasterer, T. N., Smith, T. F. & Mohr, S. C. (1998). Patterns of genome organization in bacteria. *Science*, **279**, 1827.

Frishman, D. & Mewes, H.-W. (1997). Protein structural classes in five complete genomes. *Nature Struct. Biol.* **4**, 626–628.

Gaasterland, T. & Sensen, C. W. (1996). Fully automated genome analysis that reflects user needs and preferences. A detailed introduction to the MAGPIE system architecture. *Biochimie*, **78**, 302–310.

Gelbert, L. M. & Gregg, R. E. (1997). Will genetics really revolutionize the drug discovery process? *Curr. Opin. Biotechnol.* **8**, 669–674.

Gelfand, M. S., Mironov, A. A. & Pevzner, P. A. (1996). Gene recognition via spliced sequence alignment. *Proc. Nat. Acad. Sci. USA*, **93**, 9061–9066.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H. & Oliver, S. G. (1996). Life with 6000 genes. *Science*, **274**, 546–567.

Guigo, R. (1997). Computational gene identification: an open problem. *Comput. Chem.* **21**, 215–222.

Guigo, R. & Smith, T. F. (1993). Inferring correlation between database queries: analysis of protein sequence patterns. *IEEE Trans. Patt. Analysis Mach. Intell.* **15**, 1030–1041.

Guigo, R., Johansson, A. & Smith, T. F. (1991). Automatic evaluation of protein sequence functional patterns. *Comp. Appl. Biosci.* **7**, 309–315.

Han, K. & Kim, H.-J. (1993). Prediction of common folding structures of homologous RNAs. *Nucl. Acids Res.* **21**, 1251–1257.

Henikoff, S., Pietrokovski, S. & Henikoff, J. G. (1998). Superior performance in protein homology detection with the Blocks Database servers. *Nucl. Acids Res.* **26**, 309–312.

Hertz, G. Z., Hartzell, G. W. & Stormo, G. D. (1990). Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comp. Appl. Biosci.* **6**, 81–92.

Hieter, P. & Boguski, M. (1997). Modern modifications. *Science*, **278**, 601–602.

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.-C. & Herrmann, R. (1996). Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucl. Acids Res.* **24**, 4420–4449.

Himmelreich, R., Plagens, H., Hilbert, H., Reiner, B. & Herrmann, R. (1997). Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*. *Nucl. Acids Res.* **25**, 701–712.

Hood, D. W., Deadman, M. E., Jennings, M. P., Bisercic, M., Fleischmann, R. D., Venter, J. C. & Moxon, E. (1996). DNA repeats identify novel virulence genes in Haemophilus influenzae. *Proc. Natl Acad. Sci. USA*, **93**, 11121–11125.

Humphery-Smith, I. & Blackstock, W. (1997). Proteome analysis: genomics via the output rather than the input code. *J. Protein Chem.* **16**, 537–544.

Huynen, M. A. & Bork, P. (1998). Measuring genome evolution. *Proc. Natl Acad. Sci. USA*, **95**, 5849–5856.

Huynen, M. A., Perelson, A., Vieira, W. A. & Stadler, P. F. (1996). Base pairing probabilities in a complete HIV-1 RNA. *J. Comput. Biol.* **3**, 253–274.

Huynen, M. A., Diaz-Lazcoz, Y. & Bork, P. (1997). Differential genome display. *Trends Genet.* **13**, 389–390.

Huynen, M. A., Dandekar, T. & Bork, P. (1998a). Genomics: differential genome analysis applied to the species specific features of *Helicobacter pylori*. *FEBS Letters*, **426**, 1–5.

Huynen, M. A., Doerks, T., Eisenhaber, F., Orengo, C., Sunyaev, A., Yuan, Y. & Bork, P. (1998b). Homology-based fold predictions for *Mycoplasma genitalium* proteins. *J. Mol. Biol.* **280**, 323–326.

Igarashi, T. & Kaminuma, T. (1997). Development of a cell signaling networks database. *Pac. Symp. Biocomp.* 187–197.

Jacq, B., Horn, F., Janody, F., Gompel, N., Serralbo, O., Mohr, E., Leroy, C., Bellon, B., Fasano, L., Laurenti, P. & Röder, L. (1997). GIF-DB, a WWW database on gene interactions involved in *Drosophila melanogaster* development. *Nucl. Acids Res.* **25**, 67–71.

Johnson, M. S., Srinivasan, N., Sowdhamini, R. & Blundell, T. L. (1994). Knowlegde-based protein modelling. *CRC Crit. Rev. Biochem. Mol. Biol.* **29**, 1–68.

Jurka, J., Klonowski, P., Dagman, V. & Pelton, P. (1996). CENSOR: a program for identification and elimination of repetitive elements from DNA sequences. *Comput. Chem.* **20**, 119–121.

Kahn, P. (1995). From genome to proteome: looking at a cell's proteins. *Science*, **270**, 369–370.

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., *et al.* (1996). Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis sp.* strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* **3**, 109–136.

Karp, P. D., Riley, M., Paley, S. M., Pellegrini-Toole, A. & Krummenacker, M. (1998). EcoCyc: Encyclopedia of *Escherichia coli* genes and metabolism. *Nucl. Acids Res.* **26**, 50–53.

Klenk, H. P., Clayton, R. A., Tomb, J-F., White, O., Nelson, K. E., Ketchum, K. E., Dodson, R. J., Gwinn, M., Hickey, E. K., Peterson, J. D., Richardson, D. L., Kerlavage, A. R., Graham, D. E., Kyrpides, N. C., Fleischmann, R. D., *et al.* (1997). The complete genome sequence of the hyperthermo-

philic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature*, **390**, 364–370.
Koonin, E. V. & Galperin, M. Y. (1997). Prokaryotic genomes: the emerging paradigm of genome-based microbiology. *Curr. Opin. Genet. Dev.* **7**, 757–763.
Koonin, E. V., Mushegián, A. R. & Bork, P. (1996). Non-orthologous gene displacement. *Trends Genet.* **12**, 334–336.
Koonin, E. V., Mushegian, A. R., Galperin, M. Y. & Walker, D. R. (1997). Comparison of archaeal and bacterial genomes: computer analysis of proteins sequences predicts novel functions and suggests a chimeric origin for the archaea. *Mol. Microbiol.* **25**, 619–637.
Krishna, R. & Wold, F. (1997). Identifcation of common post-translational modifications. In *Protein structure – A Practical Approach* (Creighton, T., ed.), 2nd edit., pp. 91–116, Oxford University Press.
Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V., Bertero, M. G., Bessieres, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S. C., *et al.* (1997). The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature*, **390**, 249–256.
Küster, B. & Mann, M. (1998). Identifying proteins and post-translational modifications by mass spectrometry. *Curr. Opin. Struct. Biol.* **8**, 393–400.
Liao, J. C., Hou, S.-Y. & Chao, Y.-P. (1996). Pathway analysis, engineering, and physiological considerations for redirecting central metabolism. *Biotechnol. Bioeng.* **52**, 129–140.
Link, A. J., Hays, L. G., Carmack, E. B. & Yates, J. R. (1997). Identification of the major proteome composition of *Haemophilus influenzae* type-strain NCTC 8143. *Electrophoresis*, **18**, 1314–1334.
Lukashin, A. V. & Borodovsky, M. (1998). GeneMark. hmm: new solutions for gene finding. *Nucl. Acid Res.* **26**, 1107–1115.
Lupas, A. (1997). Predicting coiled coil regions in proteins. *Curr. Opin. Struct. Biol.* **7**, 388–393.
Maniatis, T., Fritsch, E. F. & Sambrook, J. (1989). *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
Mavrovouniotis, M. L., Stephanopoulus, G. & Stephanopoulus, G. (1990). Computer-aided synthesis of biochemical pathways. *Biotechnol. Bioeng.* **36**, 1119–1132.
Medigue, C., Rouxel, T., Vigier, P., Henaut, A. & Danchin, A. (1991). Evidence for horizontal transfer in *Escherichia coli* speciation. *J. Mol. Biol.* **222**, 851–856.
Melendez-Hevia, E., Waddell, T. G., Heinrich, R. & Montero, F. (1997). Theoretical approaches to evolutionary optimization of glycolysis. *Eur. J. Biochem.* **244**, 527–543.
Mighell, A. J., Markham, A. F. & Robinson, P. A. (1997). Alu sequences. *FEBS Letters*, **417**, 1–5.
Mrazek, J. & Karlin, S. (1998). Strand compositional asymmetry in bacterial and large viral genomes. *Proc. Natl Acad. Sci. USA*, **95**, 3720–3725.
Mushegian, A. R. & Koonin, E. V. (1996a). Gene order is not conserved in bacterial evolution. *Trends Genet.* **12**, 289–290.
Mushegian, A. R. & Koonin, E. V. (1996b). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl Acad. Sci. USA*, **93**, 10268–10273.
Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10**, 1–6.
Nuño, J. C., Sánchez-Valdenebro, I., Pérez-Iratxeta, C., Meléndez-Hevia, E. & Montero, F. (1997). Network organization of cell metabolism: monosaccharide interconversion. *Biochem. J.* **324**, 103–111.
Ogata, H., Goto, S., Fujibuchi, W. & Kanehisa, M. (1998). Computation with the KEGG pathway database. *Biosystems*, **47**, 119–128.
Ogura, H., Agata, H., Xie, M., Odaka, T. & Furutani, H. (1997). A study of learning splice sites of DNA sequence by neural networks. *Comput. Biol. Med.* **27**, 67–75.
Oliver, S. G. (1997). From gene to screen with yeast. *Curr. Opin. Genet. Dev.* **7**, 405–409.
Overbeek, R., Panyushkina, E., Pronevitch, L., Selkov, E., Jr & Yunus, I. (1996). The metabolic pathway collection from EMP: the enzymes and metabolic pathways database. *Nucl. Acids Res.* **24**, 26–28.
Overbeek, R., Larsen, N., Smith, W., Maltsev, N. & Selkov, E. (1997). Representation of function: the next step. *Gene*, **191**, GC1–GC9.
Palsson, B. O. (1997). What lies beyond bioinformatics? *Nature Biotechnol.* **15**, 3–4.
Pearson, W. R. (1998). Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* **276**, 71–84.
Pedersen, A. G. & Engelbrecht, J. (1995). Investigations of *Escherichia coli* promoter sequences with artificial neural networks: new signals discovered upstream of the transcriptional startpoint. *Intelligent Systems Mol. Biol.* **3**, 292–299.
Persson, B., Zigler, J. S., Jr & Jörnvall, H. (1994). A super-family of medium-chain dehydrogenases/reductases (MDR). *Eur. J. Biochem.* **226**, 15–22.
Piatigorsky, J. & Wistow, G. (1991). The recruitment of crystallins: new functions precede gene duplication. *Science*, **252**, 1078–1079.
Pollack, J. D., Williams, M. V. & McElhaney, R. N. (1997). The comparative metabolism of the mollicutes (*Mycoplasmas*): the utility for taxonomic classification and the relationship of putative gene annotation and phylogeny to enzymatic function in the smallest free-living cells. *Crit. Rev. Microbiol.* **23**, 269–354.
Pujana, M. A., Volpini, V. & Estivill, X. (1998). Large CAG/CTG repeat templates produced by PCR, usefulness for the DIRECT method of cloning genes with CAG/CTG repeat expansions. *Nucl. Acids Res.* **26**, 1352–1353.
Rawlings, C. J. & Searls, D. B. (1997). Computational gene discovery and human disease. *Curr. Opin. Genet. Dev.* **7**, 416–423.
Richterich, P. (1998). Estimation of errors in "raw" DNA sequences: a validation study. *Genome Res.* **8**, 251–259.
Riley, M. (1998). Systems for categorizing functions of gene products. *Curr. Opin. Struct. Biol.* **8**, 388–392.
Rodriguez, R. & Vriend, G. (1997). Professional gambling. In *Proceedings of the NATO Advanced Study Institute on Biomolecular Structure and Dynamics: Recent Experimental and Theoretical Advances*, **1**, 1–10.
Rost, B. & Donoghue, S. (1997). Sisyphus and the prediction of protein structure. *Comp. Appl. Biosci.* **13**, 345–356.
Saccone, S., Caccio, S., Kusuda, J., Andreozzi, L. & Bernardi, G. (1996). Identification of the gene-richest bands in human chromosomes. *Gene*, **174**, 85–94.
Saito, K. & Nomura, M. (1994). Post-transcriptional regulation of the *str* operon in *Escherichia coli*; struc-

tural and mutational analysis of the target site for translational repressor S7. *J. Mol. Biol.* **255**, 125–139.

Sanchez, R. & Sali, A. (1997). Evaluation of comparative protein structure modeling by MODELLER-3. *Proteins: Struct. Funct. Genet.* **Suppl. 1**, 50–58.

Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56–68.

Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. (1998). SMART, a simple modular architecture research tool: Identification of signalling domains. *Proc. Natl Acad. Sci. USA*, **95**, 5857–5864.

Schuster, S. (1996). Control analysis in terms of generalized variables characterizing metabolic systems. *J. Theoret. Biol.* **182**, 259–268.

Schuster, S. & Hilgetag, C. (1994). On elementary flux modes in biochemical reaction systems at steady state. *J. Biol. Syst.* **2**, 165–182.

Schuster, S. & Hilgetag, C. (1995). What information about the conserved-moiety structure of chemical reaction systems can be derived from their stoichiometry? *J. Phys. Chem.* **99**, 8017–8023.

Selkov, E., Basmanova, S., Gaasterland, T., Goryanin, I., Gretchkin, Y., Maltsev, N., Nenashev, V., Overbeek, R., Panyushkina, E., Pronevitch, L., Selkov, E., Jr & Yunus, I. (1996). The metabolic pathway collection from EMP: the enzymes and metabolic pathways database. *Nucl. Acids Res.* **24**, 26–28.

Selkov, E., Maltsev, N., Olsen, G. J., Overbeek, R. & Whitman, W. B. (1997). A reconstruction of the metabolism of *Methanococcus janaschii* from sequence data. *Gene*, **197**, GC11–GC26.

Seressiotis, A. & Bailey, J. E. (1988). MPS: An artificially intelligent software system for the analysis and synthesis of metabolic pathways. *Biotechnol. Bioeng.* **31**, 587–602.

Serry, L. T., Nestor, P. V. & FitzGerald, G. A. (1998). Molecular evolution of the aldo-keto reductase gene superfamily. *J. Mol. Evol.* **46**, 139–146.

Smith, D. R., Doucette-Stamm, L. A., Deloughery, C., Lee, H., Dubois, J., Aldredge, T., Bashirzadeh, R., Blakely, D., Cook, R., Gilbert, K., Harrison, D., Hoang, L., Keagle, P., Lumm, P., Pothier, B., *et al.* (1997). Complete genome sequence of *Methanobacterium thermoautotrophicum* ΔH: functional analysis and comparative genomics. *J. Bacteriol.* **179**, 7135–7155.

Smith, T. F. (1998). Functional genomics - bioinformatics is ready for the challenge. *Trends Genet.* **14**, 291–293.

Smith, T. F. & Zhang, X. (1997). The challenges of genome sequence annotation or "The devil is in the details". *Nature Biotechnol.* **15**, 1222–1223.

Sonnhammer, E. L., Eddy, S. R., Birney, E., Bateman, A. & Durbin, R. (1998). Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucl. Acids Res.* **26**, 320–322.

Staden, R. (1989). Methods for discovering novel motifs in nucleic acid sequences. *Comp. Appl. Biosci.* **5**, 293–298.

Strauss, E. J. & Falkow, S. (1997). Microbial pathogenesis: genomics and beyond. *Science*, **276**, 707–712.

Tatusov, R. L., Mushegian, A. R., Bork, P., Brown, N. P., Hayes, W. S., Borodovsky, M., Rudd, K. E. & Koonin, E. V. (1996). Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. *Curr. Biol.* **6**, 279–291.

Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997). A genomic perspective on protein families. *Science*, **278**, 631–637.

Thomas, S. & Fell, D. A. (1996). Design of metabolic control for large flux changes. *J. Theoret. Biol.* **182**, 285–298.

Tiwari, S., Ramachandran, S., Bhattacharya, A., Bhattacharya, S. & Ramaswamy, R. (1997). Prediction of probable genes by Fourier analysis of genomic sequences. *Comp. Appl. Biosci.* **13**, 263–270.

Tomb, J. F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H. P., Gill, S., Dougherty, B. A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E. F., Peterson, S., *et al.* (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature*, **388**, 539–547.

Trifonov, E. N. (1996). Interfering contexts of regulatory sequence elements. *Comp. Appl. Biosci.* **12**, 423–429.

Tsukamoto, Y., Kato, J. & Ikeda, H. (1997). Silencing factors participate in DNA repair and recombination in *Saccharomyces cerevisiae*. *Nature*, **388**, 900–903.

Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. (1995). Serial analysis of gene expression. *Science*, **270**, 484–487.

Vietor, I. & Huber, L. A. (1997). In search of differentially expressed genes and proteins. *Biochim Biophys Acta*, **1359**, 187–199.

von Heijne, G. (1992). Membrane protein structure prediction, hydrophobicity analysis and the positive inside rule. *J. Mol. Biol.* **225**, 487–494.

Wirkner, U., Voss, H., Ansorge, W. & Pyerin, W. (1998). Genomic organization and promotor identification of the human protein kinase CK2 catalytic subunit alpha. *Genomics*, **48**, 71–78.

Wolfertstetter, F., Frech, K., Herrmann, G. & Werner, T. (1996). Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm. *Comp. Appl. Biosci.* **12**, 71–80.

Wootton, J. C. & Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266**, 554–571.

Xu, C. W., Mendelsohn, A. R. & Brent, R. (1997). Cells that register logical relationships among proteins. *Proc. Natl Acad. Sci. USA*, **94**, 12473–12478.

Yates, J. R., III (1998). Mass spectrometry and the age of the proteome. *J. Mass Spectrom.* **33**, 1–19.

Yuan, Y. P., Eulenstein, O., Vingron, M. & Bork, P. (1998). Towards detection of orthologues in sequence databases. *Bioinformatics*, **14**, 285–289.

Zhang, L., Zhou, W., Velculescu, V. E., Kern, S. E., Hruban, R. H., Hamilton, S. R., Vogelstein, B. & Kinzler, K. W. (1997). Gene expression profiles in normal and cancer cells. *Science*, **276**, 1268–1272.

Zweiger, G. & Scott, R. W. (1997). From expressed sequence tags to 'epigenomics': an understanding of disease processes. *Curr. Opin. Biotechnol.* **8**, 684–687.

Edited by P. E. Wright

(Received 22 May 1998; received in revised form 13 August 1998; accepted 13 August 1998)

Applied and Environmental Microbiology, Aug. 1997, p. 3151–3157 Vol. 63, No. 8
0099-2240/97/$04.00+0

Characterization of the Gene Encoding an Extracellular Laccase of *Myceliophthora thermophila* and Analysis of the Recombinant Enzyme Expressed in *Aspergillus oryzae*

RANDY M. BERKA,[1]* PALLE SCHNEIDER,[2] ELIZABETH J. GOLIGHTLY,[1] STEPHEN H. BROWN,[1] MARK MADDEN,[1] KIMBERLY M. BROWN,[1] TORBEN HALKIER,[2] KRISTINE MONDORF,[2] AND FENG XU[1]

Novo Nordisk Biotech, Inc., Davis, California,[1] and Novo Nordisk A/S, Bagsværd, Denmark[2]

Received 31 January 1997/Accepted 20 May 1997

A genomic DNA segment encoding an extracellular laccase was isolated from the thermophilic fungus *Myceliophthora thermophila*, and the nucleotide sequence of this gene was determined. The deduced amino acid sequence of *M. thermophila* laccase (MtL) shows homology to laccases from diverse fungal genera. A vector containing the *M. thermophila* laccase coding region, under transcriptional control of an *Aspergillus oryzae* α-amylase gene promoter and terminator, was constructed for heterologous expression in *A. oryzae*. The recombinant laccase expressed in *A. oryzae* was purified to electrophoretic homogeneity by anion-exchange chromatography. Amino-terminal sequence data suggests that MtL is synthesized as a preproenzyme. The molecular mass was estimated to be approximately 100 to 140 kDa by gel filtration on Sephacryl S-300 and to be 85 kDa by sodium dodecyl sulfate-polyacrylamide gel electrophoresis. Carbohydrate analysis revealed that MtL contains 40 to 60% glycosylation. The laccase shows an absorbance spectrum that is typical of blue copper oxidases, with maxima at 276 and 589 nm, and contains 3.9 copper atoms per subunit. With syringaldazine as a substrate, MtL has optimal activity at pH 6.5 and retains nearly 100% of its activity when incubated at 60°C for 20 min. This is the first report of the cloning and heterologous expression of a thermostable laccase.

Laccases (EC 1.10.3.1) are multicopper enzymes that catalyze the oxidation of a variety of phenolic compounds, with concomitant reduction of O_2 to H_2O. These polyphenol oxidases are widely distributed among plant (9, 10, 41, 51) and fungal (8, 14) species; however, their biological significance is unclear. Among the filamentous fungi, approximately 30 laccases have been identified in various organisms. These laccases may be involved in conidial pigmentation (4, 22), lignin degradation (2, 17, 34, 39, 56, 57, 63), pathogenicity (5), and formation of fruiting bodies (38). Interest in laccases has been fueled by their potential uses in detoxification of environmental pollutants (13, 15, 16, 26, 46, 54, 61), prevention of wine decoloration (37), paper processing (50), enzymatic conversion of chemical intermediates (1), and production of useful chemicals from lignin (57).

For any of these potential applications to become a reality, an inexpensive source of laccase must be obtained. In most fungi, laccases are produced at levels that are too low for commercial purposes. Cloning of the laccase genes followed by heterologous expression may provide higher enzyme yields. A number of genes encoding fungal laccases are cloned, including those from basidiomycetes such as *Trametes* (*Coriolus*) *versicolor* (32, 33), *Trametes villosa* (75), *Coriolus hirsutus* (35), *Rhizoctonia solani* (69), *Agaricus bisporus* (49), *Phlebia radiata* (56), basidiomycete PM1 (23), and ascomycetes *Cryphonectria parasitica* (19), *Aspergillus nidulans* (4), *Podospora anserina* (28), and *Neurospora crassa* (29). Collectively, the amino acid sequences deduced from these genes suggest that the overall structure of fungal laccases is similar to that of ascorbate oxidase from *Zucchini* (43).

The laccase genes from *C. hirsutus* and *P. radiata* have been expressed in *Saccharomyces cerevisiae* (35) and *Trichoderma reesei* (55), respectively. The yeast *GAL10* promoter was used to direct the expression of enzymatically active *C. hirsutus*, with yields of approximately 5 mg per liter (74). The *P. radiata* laccase was secreted at a level of about 20 mg per liter by using the promoter and terminator regions of the *T. reesei cbh1* gene (55). Clearly, higher enzyme titers are required for commercial enzyme production. Several *Aspergillus* species, including *Aspergillus oryzae*, are well-established as good expression systems for the heterologous production of industrial enzymes (7, 20, 24). We hypothesized that *A. oryzae* might also be well suited as a host for laccase expression and secretion.

It is well documented that thermophilic fungi may comprise a rich source of thermostable industrial enzymes (53). Furthermore, thermal tolerance is an attractive feature for many biotechnological applications of enzymes. The thermophilic fungus *Myceliophthora thermophila* (telomorph = *Thielavia heterothallica*) was described previously as a producer of cellulase and xylanase enzymes with pronounced thermal resistance (48, 53, 58–60, 76). *M. thermophila* was first described by Apinis (3) and given the name *Sporotrichum thermophile*. Its taxonomic position was reassigned to the genus *Chrysosporium* (68) and later to its current genus (65). Our objectives were to determine if *M. thermophila* produced a thermostable extracellular laccase, clone the gene encoding it, express this gene in *A. oryzae*, and biochemically characterize the resulting enzyme.

MATERIALS AND METHODS

Fungal strains and plasmids. Genomic DNA was isolated from *M. thermophila* CBS 117.65. *Escherichia coli* JM101 (45) was used for construction and routine propagation of laccase expression vectors. The fungal host for laccase expression was a uridine-requiring (*pyrG*) mutant of the α-amylase-deficient *A. oryzae* strain HowB104.

The vector pMWR3 was constructed by inserting the *A. oryzae* α-amylase promoter and terminator elements from pTAKA17 (11, 21) into pUC18 (72). In this vector, there are a *SwaI* site at the end of the promoter and an *NsiI* site at the beginning of the terminator for directional cloning. The cloning vehicle

* Corresponding author. Mailing address: Novo Nordisk Biotech, Inc., 1445 Drew Ave., Davis, CA 95616. Phone: (916) 757-4974. Fax: (916) 758-0317. E-mail: ramb@novo.dk.

pUC519 (6a) was derived by inserting a small linker containing *Nsi*I, *Cla*I, *Xho*I, and *Bgl*II restriction sites between the adjacent *Bam*HI and *Xba*I sites of pUC119 (66).

Materials. The chemicals, buffers, and substrates used were commercial products of at least reagent grade. Endo/*N*-glycosidase F and pyroglutamate aminopeptidase were purchased from Boehringer Mannheim (Indianapolis, Ind.). Chromatography was done by using fast protein liquid chromatography system (Pharmacia LKB, Uppsala, Sweden) or a conventional open low-pressure system. Spectroscopic assays were conducted with either a UV160 UV-visible light spectrophotometer (Shimadzu, Inc., Columbia, Md.) or a microplate reader (Molecular Devices, Menlo Park, Calif.).

DNA extraction and hybridization analysis. Total cellular DNA was extracted from *M. thermophila* cells by the procedure described by Timberlake and Barnard (64). Genomic DNA samples were analyzed by Southern hybridization (25) under conditions of mild stringency (i.e., 5× SSPE [1× SSPE is 0.18 M NaCl, 10 mM NaH_2PO_4, and 1 mM EDTA; pH 7.7], 35% formamide, 0.3% sodium dodecyl sulfate [SDS]). The laccase-specific probe fragment (approximately 1.5 kb) comprised the 5′ portion of the *N. crassa lcc-1* gene. The purified probe fragment was radiolabeled by nick translation (42) with [α-^{32}P]dCTP (Amersham) and added to the hybridization buffer at an activity of approximately 10^6 cpm per ml. The mixture was incubated overnight at 45°C. Following incubation, the membrane filters were washed once in 0.2× SSPE with 0.1% SDS at 45°C and then twice in 0.2× SSPE (no SDS) at the same temperature. The filters dried on paper towels for 15 min and then were wrapped in Saran Wrap and exposed to X-ray film overnight at −70°C with intensifying screens.

DNA libraries and identification of laccase clones. A genomic DNA library was constructed in λ-EMBL4 (62). Briefly, DNA was partially digested with *Sau*3AI and size fractionated on low-melting-point agarose gels. DNA fragments migrating between 9 and 23 kb were excised and eluted from the gel by using β-agarase (New England Biolabs, Beverly, Mass.). The eluted DNA fragments were ligated with *Bam*HI-cleaved and dephosphorylated λ-EMBL4 vector arms, and the ligation mixtures were packaged by using commercial packaging extracts (Stratagene, La Jolla, Calif.). The packaged DNA library was plated and amplified on *E. coli* K802 cells (42). Approximately 20,000 plaques were screened by plaque hybridization (25) with the radiolabeled *N. crassa* laccase gene fragment under the conditions described above. Plaques which gave hybridization signals with the probe were purified twice on *E. coli* K802 cells, and DNA from three of these phage was purified by using a Qiagen Lambda kit (Qiagen, Inc., Chatsworth, Calif.).

Analysis of laccase genes. Restriction mapping was completed by standard methods (40). DNA sequencing was done with a model 373A automated DNA sequencer (Applied Biosystems, Inc., Foster City, Calif.) by using the primer walking technique with dye-terminator chemistry (30). The final nucleotide sequence was determined on both strands. Oligonucleotides were synthesized on an Applied Biosystems model 394 DNA/RNA synthesizer.

Construction of laccase expression vectors. Construction of the laccase expression vector pRaMB5 is outlined in Fig. 1. The promoter directing transcription of the laccase gene segment was obtained from the *A. oryzae* α-amylase (TAKA-amylase) gene (21). The α-amylase polyadenylation/transcription terminator region from pTAKA17 was also used in construction of this vector (21).

Cotransformation of *A. oryzae*. Methods for cotransformation of *A. oryzae* were described by Christensen et al. (21). Equal amounts (approximately 5 μg each) of laccase vector and one of the following plasmids were used: p*pyrG* (Fungal Genetics Stock Center, Kansas City, Kans.), which contains the *A. nidulans pyrG* gene (47), or pSO2, which harbors the *A. oryzae pyrG* gene. Prototrophic (Pyr^+) transformants were selected on *Aspergillus* minimal medium (52), and the transformants were screened for the ability to produce laccase on minimal medium containing 1 mM 2,2′-azinobis(3-ethylbenzthiazolin-6-sulfonic acid) (ABTS). Cells that secreted active laccase oxidized the ABTS, producing a green halo surrounding the colony.

Analysis of laccase-producing transformants. Transformants that produced laccase activity on agar plates were purified twice through conidiospores, and spore suspensions in sterile 0.01% Tween 80 were made from each. The density of spores in each suspension was estimated spectrophotometrically by absorption at 595 nm. Approximately 0.5 absorbance units of spores was used to inoculate 25 ml of shake flask medium in 125-ml plastic flasks. The shake flask medium contained the following (per liter): 1 g of $CaCl_2 \cdot 2H_2O$, 2 g of yeast extract, 1 g of $MgSO_4$, 2 g of citric acid, 5 g of KH_2PO_4, 1 g of urea, 2 g of $(NH_4)_2SO_4$, 20 g of maltodextrin, and 0.5 ml of trace elements solution (36). The cultures were incubated at 37°C with vigorous aeration (approximately 200 rpm) for 4 to 5 days. Culture broths were harvested by centrifugation, and the amount of laccase activity in the supernatant was determined. Transformants producing the highest levels of the recombinant *M. thermophila* laccase (r-MtL) in shake flask cultures were also grown in laboratory fermentors.

Laccase assays. The syringaldazine oxidase activity of r-MtL was determined by using 19 μM syringaldazine and monitoring the absorbance change at 530 nm (extinction coefficient = 65 mM^{-1} cm^{-1} [6]). One syringaldazine oxidation unit (SOU) was defined as the amount of enzyme that oxidizes 1 μmol of substrate per min in 1 ml at 20°C. ABTS oxidation assays were done by using 1 mM ABTS and monitoring the absorbance change at 418 or 405 nm (extinction coefficient = 36 or 35 mM^{-1} cm^{-1}, respectively [18]). Britton and Robinson (B&R) buffers, made by mixing 0.1 M boric acid–0.1 M acetic acid–0.1 M phosphoric acid with 0.5 M NaOH to the desired pH, were used to determine the pH activity profile of r-MtL. Thermostability analysis of r-MtL was performed by using 0.8 to 1.2 μM samples preincubated in B&R buffer (pH 6) at various temperatures. The samples were assayed for syringaldazine oxidase activity after a 430-fold dilution at room temperature.

Purification of native MtL from *M. thermophila* culture broth. *M. thermophila* was grown for 5 days at 42°C in medium which contained the following: 1% glucose, 4% dextrin, 2% ammonium citrate, 0.1% $MgSO_4 \cdot 7H_2O$, 1% KH_2PO_4, 0.1% $CaCl_2$, 0.01% $FeSO_4 \cdot 7H_2O$, 0.05% $CuSO_4 \cdot 5H_2O$, 0.01% pluronic antifoam, and 0.5% PWH salts [containing (per liter) 0.3 g of $ZnCl_2$, 0.6 g of $FeSO_4 \cdot 7H_2O$, 0.25 g of $CuSO_4$, 0.35 g of $MnSO_4 \cdot H_2O$, 0.2 g of $(NH_4)_6Mo_7O_{24} \cdot 4H_2O$, and 6 g of tetrasodium-EDTA]. The mycelia were removed by filtration, and the culture broth was washed, filtered, adjusted to pH 7, and applied to a Q-Sepharose (Hiload 26/10; Pharmacia) column that was preequilibrated with 0.1 M phosphate buffer, pH 7. The laccase concentration of the crude broth was approximately 5 mg per liter. The laccase activity eluted with a gradient of 0 to 1 M NaCl. An 11-fold purification and recovery yield of 56% were achieved.

Purification of r-MtL from *A. oryzae* culture broth. A washed, concentrated broth sample (pH 7.6; conductivity = 0.8 mS) was loaded onto a Q-Sepharose XK26 column (120 ml; Pharmacia) preequilibrated with 10 mM Tris, pH 7.5. MtL has an intense blue color (corresponding to an absorbance peak at 600 nm) that is typical of multicopper oxidases (44). One group of blue fractions was collected after the column was loaded and washed. A second group eluted with a linear gradient of 0 to 2 M NaCl. SDS-polyacrylamide gel electrophoresis (SDS-PAGE) analysis showed that this preparation was essentially pure laccase. A purification of 121-fold and recovery of 67% were achieved. The purified r-MtL showed no activity loss over a 5-week-long storage frozen in Q-Sepharose elution buffer at −20°C.

Analyses of amino acid composition, carbohydrate content, N-terminal and C-terminal sequences, copper content, and native molecular mass. N-terminal sequencing was done with an Applied Biosystems model 476A protein sequencer. The sequencing reagents were from Perkin-Elmer/Applied Biosystems Division (Foster City, Calif.). A 1090L high-pressure liquid chromatography system (Hewlett-Packard Co., Wilmington, Del.) equipped with diode array detection at 215 and 280 nm and 3D Chemstation software was used for the separation of CNBr- and protease-generated enzyme fragments. Separations were done on a Vydac C_4 or C_{18} reverse-phase column (Vydac, Hesperia, Calif.). C-terminal sequencing was done by J. M. Bailey of Hewlett-Packard Co. Total amino acid analysis, from which the extinction coefficient of r-MtL was determined, was done with a Hewlett-Packard 1090 AminoQuant instrument.

Hydrolyses of protein-bound carbohydrate for monosaccharide compositional analysis were done in duplicate. Lyophilized samples were hydrolyzed in evacuated sealed glass tubes with 100 μl of 2 M trifluoroacetic acid (TFA) for 1 and 4 h at 100°C. Monosaccharides were separated by high-performance anion-exchange chromatography using a CarboPac PA1 column (Dionex Corporation, Sunnyvale, Calif.), eluted with 16 mM NaOH, and detected by pulsed amperometric detection. Due to the different stability and release of the monosaccharides in 2 M TFA, the amounts of glucosamine and mannose were determined after 4 h of hydrolysis, whereas the amount of galactose was determined after 1 h of hydrolysis. Deglycosylation was also achieved by using endo/*N*-glycosidase F (Boehringer Mannheim) according to the manufacturer's instructions, and the carbohydrate content of r-MtL was estimated from the mobility difference in SDS-PAGE. Enzymatic removal of the N-terminal pyroglutamate residue was done with pyroglutamate aminopeptidase (Boehringer Mannheim) in accordance with the manufacturer's instructions. About 80 μg of r-MtL was treated with 4 μg of peptidase with or without 1 M urea or 0.1 M guanidine HCl and then transferred to a polyvinylidene difluoride membrane for sequencing. About 20 pmol of peptidase-treated protein was obtained and sequenced.

SDS-PAGE and native isoelectric focusing (IEF) analysis were done on commercial apparatus (Novex, San Diego, Calif., and Bio-Rad Laboratories, Hercules, Calif.). Proteins were stained with Coomassie brilliant blue. Gel filtration analyses were done on a Sephacryl S-300 (Pharmacia) column, and the native molecular mass was estimated by using blue dextran (2,000 kDa), bovine immunoglobulin G (158 kDa), bovine serum albumin (66 kDa), ovalbumin (45 kDa), and horse heart myoglobin (17 kDa) to calibrate the column.

The copper content was determined by the photometric titration method of Felsenfeld (27) and by atomic absorption spectroscopy.

The extinction coefficient for r-MtL was calculated on the basis of amino acid analysis, and the molecular mass was deduced from the DNA sequence.

Nucleotide sequence accession number. The nucleotide sequence of the *lcc1* coding region was determined and deposited in the GenSeq database under accession no. T10922.

RESULTS

Cloning and characterization of sequence of the laccase gene from *M. thermophila*. Genomic DNA was prepared from *N. crassa* and *M. thermophila*, digested with *Bam*HI, fractionated by agarose gel electrophoresis, blotted, and probed under conditions of mild stringency with a radiolabeled fragment

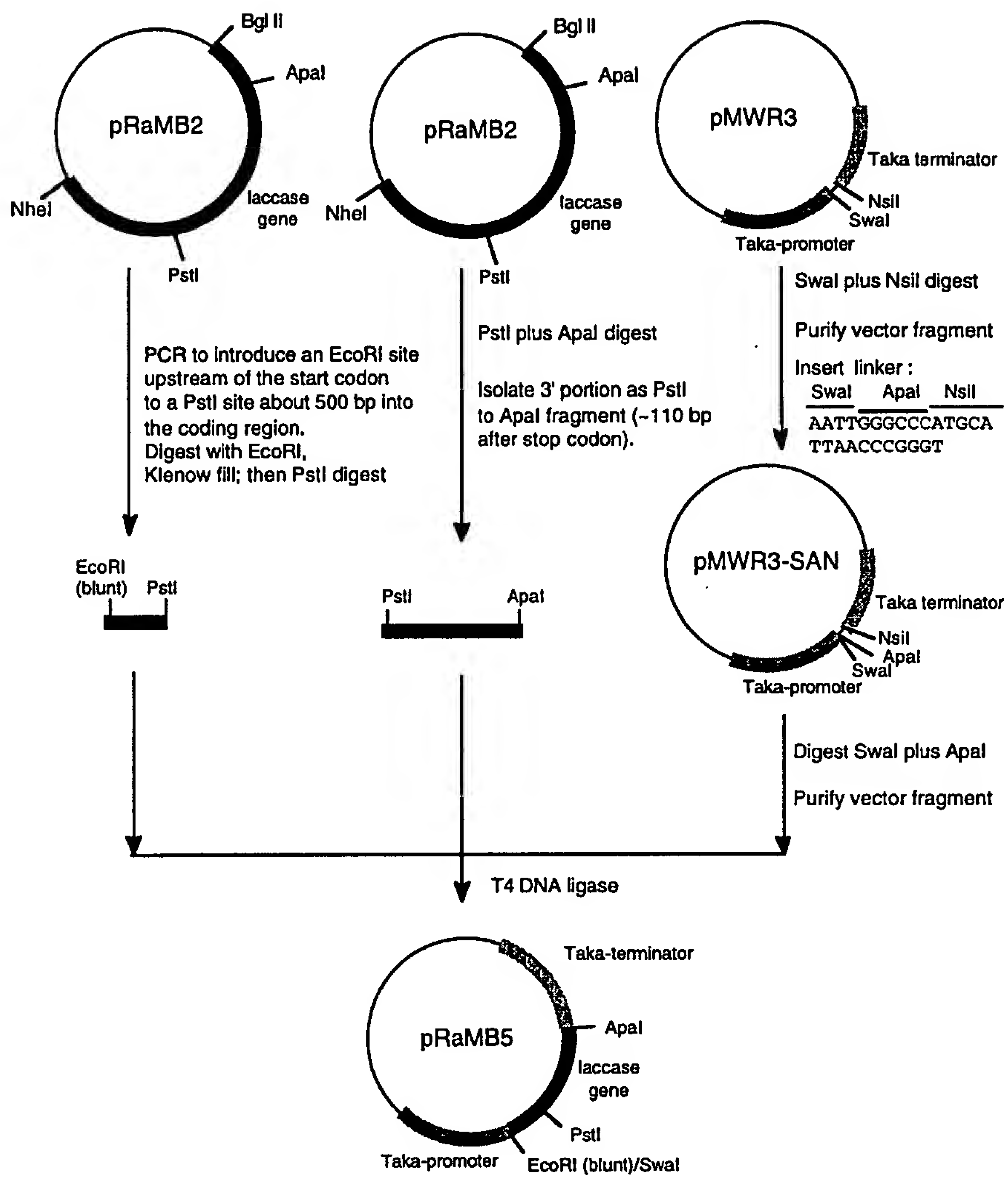


FIG. 1. Scheme for construction of the vector pRaMB5 for *Myceliophthora* laccase expression in *Aspergillus*. First, plasmid pMWR3 was modified by inserting a small linker that contained an *Apa*I site between the *Swa*I and *Nsi*I sites, creating a plasmid called pMWR3-SAN. Second, *Pfu* polymerase-directed PCR (Stratagene) was used to amplify a short DNA segment encoding the 5′ portion of MtL, from the start codon to an internal *Pst*I site (approximately 0.5 kb). The forward primer for this PCR was designed to create an *Eco*RI site just upstream of the start codon (Fig. 2). Next, the amplified fragment was digested with *Eco*RI and *Pst*I and subcloned into an M13mp18 sequencing vector, and its nucleotide sequence was verified. This fragment was subsequently excised by cleavage with *Eco*RI and *Pst*I (during this step, the *Eco*RI site was made blunt by treatment with deoxynucleoside triphosphates and DNA polymerase I [Klenow fragment]) and purified by agarose gel electrophoresis. The 3′ portion of the *lcc1* coding region was excised from pRaMB2 as a 2-kb *Pst*I-*Apa*I fragment (this segment also contains approximately 110 bp from the 3′ untranslated region). Lastly, these two fragments were combined with *Swa*I- and *Apa*I-cleaved pMWR3-SAN in a three-part ligation reaction to generate the laccase expression vector pRaMB5.

encoding a portion of the *N. crassa* laccase gene. A single laccase-specific DNA fragment was detected in both genomic digests. We then screened approximately 20,000 plaques from an *M. thermophila* genomic DNA library. Eight plaques that hybridized strongly to the probe were identified. DNA was isolated from three of these plaques, cleaved with *Eco*RI, and analyzed by agarose gel electrophoresis and Southern hybridization. All three clones contained a 7.5-kb *Eco*RI fragment which hybridized to the laccase-specific probe. One fragment was subcloned into pBR322 (12) to generate plasmid pRaMB1. The entire *M. thermophila* laccase gene (*lcc1*) coding region was contained within a 3.2-kb *Nhe*I-*Bgl*II segment that was subcloned into pUC119 (66) to give plasmid pRaMB2. The nucleotide sequence of this segment was determined on both strands by the primer walking method (30).

The positions of six introns (85, 84, 102, 72, 147, and 95 nucleotides in length) within the *lcc1* coding region were determined by comparing the deduced amino acid sequence of MtL to that of *N. crassa* laccase and by applying the consensus rules for intron features in filamentous fungi (31). Additionally, the amino acid sequences of several internal peptide fragments from recombinant MtL were determined, and the correct reading frame for the *lcc1* gene as well as the positions of the second, third, and sixth introns was verified. The 1,860 nucleotides of coding sequence are 65.5% G+C, with a strong bias (90%) for codons ending in G or C.

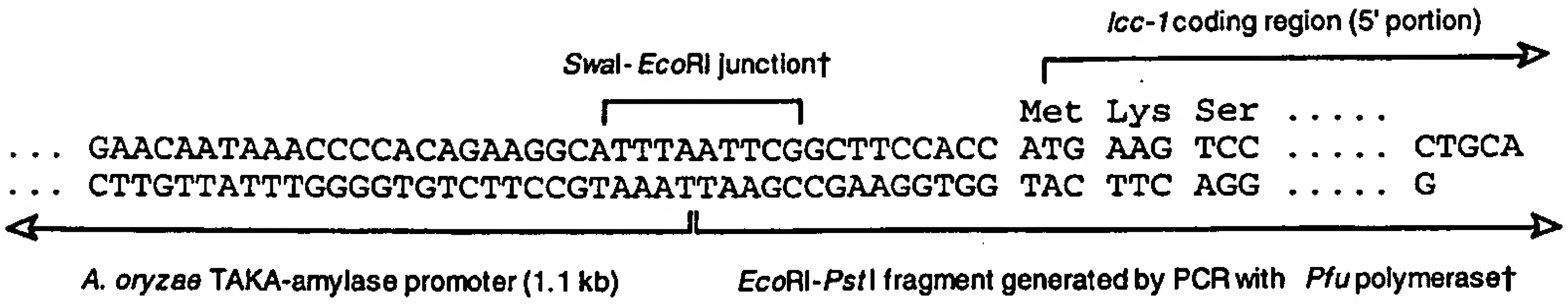


FIG. 2. Scheme for joining an *A. oryzae* α-amylase promoter to the MtL coding region in the expression vector pRaMB5. †, *Eco*RI site that was made blunt by treatment with DNA polymerase I (Klenow fragment) and deoxynucleoside triphosphates.

The deduced amino acid sequence of MtL shares identity with laccases of the following species: *Podospora anserina* (65%), *N. crassa* (60%), *C. parasitica* (53%), *Agaricus bisporus* (24%), *P. radiata* (22%), *Pleurotus ostreatus* (22%), *R. solani* (20%), *T. versicolor* (22%), *T. villosa* (22%), and *A. nidulans* (15%). Similarity is highest in the regions that correspond to the four histidines and one cysteine that form the trinuclear copper cluster (23, 44, 49). There are 11 potential sites (Asn-X-Ser/Thr) for N-linked glycosylation in the deduced amino acid sequence of MtL.

The first 22 amino acids in the deduced primary structure of MtL appear to comprise a canonical signal peptide with a predicted cleavage following an Ala residue (67). The purified extracellular forms of both native MtL and r-MtL are blocked with N-terminal pyroglutamate residues. Enzymatic removal of these residues followed by amino acid sequencing suggests that mature MtL begins with a Gln residue. Thus, MtL is apparently synthesized as a 619-residue preproenzyme having a 22-residue signal peptide and a propeptide of 25 residues.

Expression of *Myceliophthora* laccase. The expression vector pRaMB5 (Fig. 1 and 2) was used to generate *A. oryzae* cotransformants which produce r-MtL that were detected by incorporation of ABTS into selective media. As determined by using the *pyrG* gene from *A. nidulans* or *A. oryzae* as the selectable marker, the frequencies of laccase-producing cotransformants among Pyr^+ colonies were 59% with *A. nidulans pyrG* as the selected marker and 31% with *A. oryzae pyrG* as the marker. Several cotransformants that produced intense color reactions on ABTS plates were grown in shake flask cultures to quantitate the amount of r-MtL produced. The amount of extracellular laccase activity produced ranged from 0.49 to 0.85 SOU/ml (Table 1). On the basis of the specific activity of 45 SOU/mg (see below), the level of r-MtL secreted in these shake flask cultures ranged from 11 to 19 mg per liter. Preliminary SDS-PAGE analyses of culture broth samples showed a prominent laccase band at approximately 85 kDa, which is similar to the size of the native enzyme purified from *M. thermophila*.

TABLE 1. MtL expression among selected *A. oryzae* transformants[a]

Transformant	Transforming DNAs[b]	SOU/ml in shake flask
Control	None	0.00
RaMB5.15	pRaMB5 + p*pyrG*	0.85
RaMB5.30	pRaMB5 + p*pyrG*	0.71
RaMB5.33	pRaMB5 + p*pyrG*	0.60
RaMB5.108	pRaMB5 + pSO2	0.68
RaMB5.111	pRaMB5 + pSO2	0.70
RaMB5.121	pRaMB5 + pSO2	0.49
RaMB5.142	pRaMB5 + pSO2	0.54

[a] *A. oryzae* HowB104 *pyrG* was the host strain.
[b] Plasmids p*pyrG* and pSO2 contain the *pyrG* genes of *A. nidulans* and *A. oryzae*, respectively.

Biochemical characterization of r-MtL produced in *A. oryzae*. Q-Sepharose chromatography yielded two active fractions that contained essentially pure laccase (one passed through the column, and another was eluted by a 0 to 2 M NaCl gradient). Purified r-MtL has a molecular mass of 100 to 140 kDa as determined by S-300 gel filtration (data not shown) and a molecular mass of 75 to 95 kDa as determined by SDS-PAGE (Fig. 3). Under nondenaturing conditions, both r-MtL fractions had a pI of 4.2. Treatment of the purified enzyme with *N*-glycosidase resulted in a decrease in the apparent molecular mass to approximately 73 kDa (Fig. 3). The increased mobility on SDS-PAGE after deglycosylation suggested that N-linked carbohydrates accounted for approximately 14% of the total mass of each subunit. Total-carbohydrate analysis showed that the laccase fractions that passed through the Q-Sepharose column (preequilibrated with 10 mM Tris, pH 7.5) contained 26 mol of glucosamine, 67 mol of galactose, 9 mol of glucose, and 138 mol of mannose per mol of enzyme. The laccase fractions that eluted from the Q-Sepharose with NaCl had 23 mol of glucosamine, 38 mol of galactose, 4 mol of glucose and 85 mol of mannose per mol of enzyme.

Attempts to directly sequence the N terminus of r-MtL from samples either in desalted solution or on polyvinylidene difluoride membranes were unsuccessful. Treatment of r-MtL with pyroglutamate aminopeptidase yielded a protein with a deblocked N terminus, beginning 48 residues after the putative translation start (Met). Sequencing of internal peptides generated by CNBr cleavage confirmed the DNA sequence and several intron and exon assignments. Direct C-terminal sequencing indicated that r-MtL had a C terminus of -Gly-Leu.

The UV-visible absorbance spectrum of r-MtL shows absorption maxima at 276 and 589 nm. The ratio of the absorbance at 280 nm to the absorbance at 600 nm was 35. This is higher than reported for *T. villosa* (75) and *R. solani* (69) laccases, suggesting that MtL contains more tryptophan, phe-

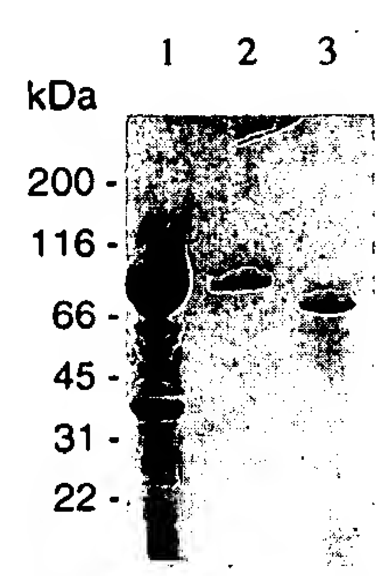


FIG. 3. SDS-PAGE profile for the purification of r-MtL. Lane 1, concentrated *A. oryzae* culture broth to be loaded onto Q-Sepharose; lane 2, purified r-MtL; lane 3, r-MtL treated with endo-/*N*-glycosidase F. The gel was stained with Coomassie brilliant blue.

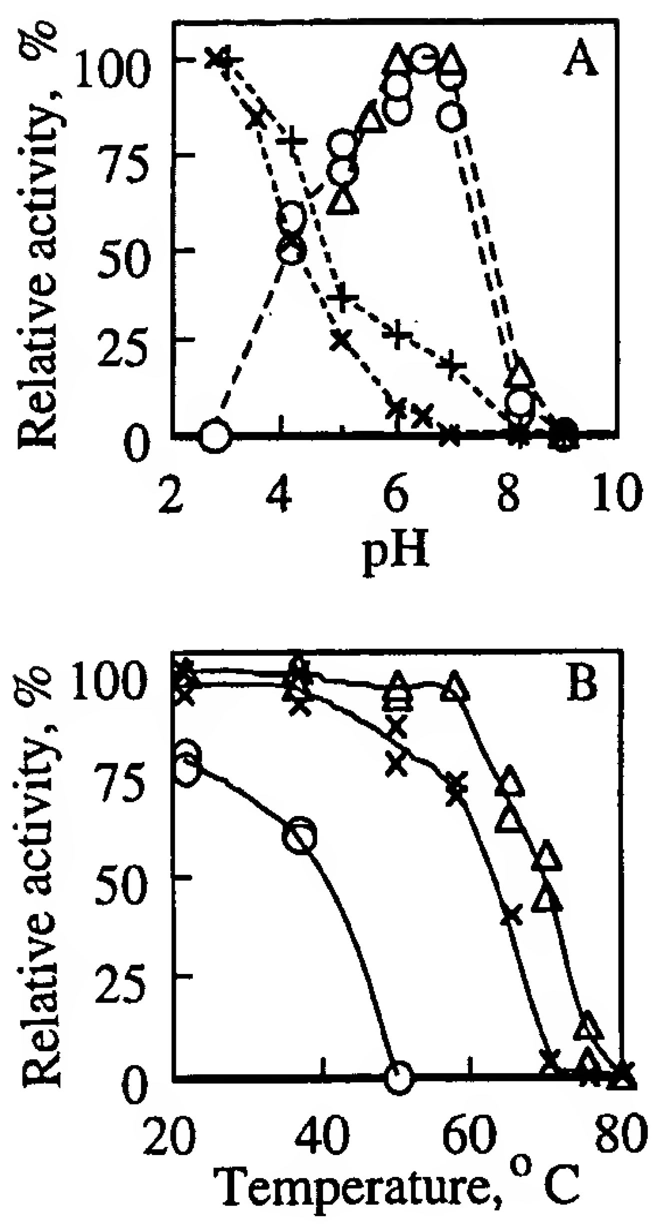


FIG. 4. Dependence of r-MtL activity on pH and temperature. (A) Laccase activity as a function of pH (normalized to the optimum activity value) of native (nonrecombinant) MtL with ABTS as a substrate (+), r-MtL with ABTS as a substrate (×), native MtL with syringaldazine as a substrate (△), and r-MtL with syringaldazine as a substrate (○). (B) Thermostability of r-MtL. Enzyme samples (0.8 to 1.2 μM) were preincubated in B&R buffer (pH 6) for 0.3 (△), 1.0 (×), and 17 (○) h at various temperatures, diluted 430-fold, and assayed for residual activity in B&R buffer (pH 6) at 20°C with syringaldazine as the substrate. The activities were normalized to the initial activity at 20°C before the preincubation.

nylalanine, and cysteine. This suggestion was confirmed by comparing the amino acid compositions deduced from the corresponding gene sequences. Photometric titration and atomic absorption spectroscopy indicated a stoichiometry of 3.9 copper atoms per enzyme subunit.

r-MtL had an optimal pH of 6.5 with syringaldazine as the substrate (Fig. 4A). At the optimum pH, the specific activity of r-MtL with syringaldazine used as a substrate was 45 SOU per mg. With ABTS as a substrate, r-MtL showed maximum activity at the lowest pH studied (pH 2.7). The difference in optimum pH values for ABTS and syringaldazine substrates is consistent with the hypothesis that electron transfer kinetics are more important than substrate binding in determining the pH activity profile of laccase (71). Thermostability analysis shown in Fig. 4B indicated that the upper temperature limit for retaining full activity after a 20-min preincubation was 60°C.

Purification and characterization of native MtL from *M. thermophila*. Q-Sepharose chromatography yielded an 11-fold purification of laccase from *M. thermophila*. Purified MtL migrated as a diffuse band in SDS-PAGE with a molecular mass of 80 kDa, and on IEF gels it had an isoelectric point of 4.2. MtL showed a UV-visible spectrum with an absorbance maximum of 280 and two smaller shoulders at 330 and 600 nm. The weak absorbance maximum at 600 nm indicated the presence of apo-MtL (copper depleted) in the preparation, possibly resulting from instability during purification. The pH activity profile of native MtL on syringaldazine is also shown in Fig. 4A. At optimum pH, MtL had an activity of 15 SOU per mg. N-terminal sequencing analysis indicated that the amino terminus was also blocked.

DISCUSSION

We cloned a genomic DNA segment encoding an extracellular laccase from the thermophilic fungus *M. thermophila*. On the basis of a comparison of its deduced amino acid sequence, MtL shows identity with laccases from diverse fungal genera. However, the greatest degree of sequence identity (53 to 65%) is between MtL and the laccases of related species in the order Sphaeriales, such as *P. anserina* (28), *N. crassa* (29), and *C. parasitica* (19). Laccases from basidiomycetes show more-limited sequence identity (20 to 24%) to MtL. Interestingly, the conidial laccase (*yA* gene product) of *A. nidulans* (4) shows the lowest degree of similarity, suggesting a possible evolutionary or functional difference between conidial and secreted laccases.

At the genomic level, the *lcc1* gene of *M. thermophila* also has the highest homology with laccase genes from *P. anserina*, *N. crassa*, and *C. parasitica*; however, their architecture is very different. For example, *N. crassa lcc1* contains a single intron, *lac2* from *P. anserina* has three introns, *lcc1* from *M. thermophila* laccase has six intervening sequences, and the *C. parasitica* laccase gene has 12 introns. The position of the first intron is conserved among laccase genes from *M. thermophila*, *N. crassa*, and *P. anserina*. Additionally, introns II and III in *P. anserina lac2* align with the third and fourth introns of *M. thermophila lcc1*. The positions of five intervening sequences in *M. thermophila lcc1* are conserved in the *C. parasitica* laccase gene. Therefore, we postulate that the laccase genes from these four species were derived from a common ancestral form but diverged during evolution. The comparatively low level of sequence similarity between the laccase genes of basidiomycetes and ascomycetes probably reflects the large phylogenetic distance between these fungal classes.

The primary structures of the laccase gene products from *Neurospora*, *Podospora*, and *Myceliophthora* predict similar mechanisms of posttranslational processing. On the basis of the rules of von Heijne (67), the predicted signal peptide cleavage site for MtL lies after the first 22 amino acids. However, direct sequencing of the amino terminus of native and recombinant forms of MtL suggested that the first residue of the mature enzyme is Gln_{49}. Therefore, residues 23 through 48 probably comprise a propeptide whose proteolytic removal occurs during maturation of MtL, leaving Gln_{49} as the first amino acid residue of the mature enzyme, which may subsequently cyclize to pyroglutamate, yielding a blocked N terminus. It was reported that *N. crassa* and *P. anserina* laccases are processed similarly at their amino-terminal ends (28, 29). In addition, *N. crassa* laccase is also reportedly processed at its C terminus, resulting in the proteolytic removal of 13 residues (29). The processing site is contained within the sequence Asp-Ser-Gly-Leu ↓ Arg_{558} (where ↓ designates the cleavage site). Strikingly similar sequences exist near the C termini of MtL (Asp-Ser-Gly-Leu-Lys_{560}) and *P. anserina* laccase (Asp-Ser-Gly-Leu-Lys_{559}). C-terminal sequencing showed that the C terminus of MtL was Gly-Leu, indicating that the enzyme was processed (Asp-Ser-Gly-Leu ↓ Lys_{560}) similarly to *N. crassa* laccase and 13 residues were removed. C-terminal processing of *P. anserina* laccase was also postulated (28). It is particularly interesting that the *C. parasitica* laccase has Asp-Ser-Gly-Val as its C terminus, and thus no processing may be needed. The importance of C-terminal processing for the catalytic activity of these enzymes is unknown, and the protease involved in this cleavage has not been identified.

By ion-exchange chromatography, r-MtL from *A. oryzae* fermentor broth was separated into multiple isoforms with different elution properties. However, no significant difference among these isoforms was seen in terms of SDS-PAGE, native PAGE, native IEF, S-300 gel filtration, UV-visible spectrum, specific activity towards syringaldazine, and unblocked-N-terminus sequencing measurements. On the basis of total-carbohydrate analyses, it appears likely that the different elution patterns of various r-MtL isoforms from Q-Sepharose arose from differential glycosylation with galactose and mannose. Total-carbohydrate analyses also gave an estimate of 33 to 60% total glycosylation, of which 14% is estimated to be N linked, on the basis of the mobility change on SDS-PAGE after *N*-glycosidase treatment. The r-MtL differed from native MtL in two other respects. First, the molecular mass of native MtL (80 kDa) was less than that of r-MtL (85 kDa), presumably reflecting differences in glycosylation. Second, the specific activity of native MtL (15 SOU/mg) was lower than that of r-MtL (45 SOU/mg). Since native MtL also had a lower absorbance at 600 nm, the decreased specific activity is probably due to a percentage of holoenzyme lower than that of r-MtL. Since the type II copper is easily depleted (73), extra copper ions were added to the culture medium of *A. oryzae* transformants expressing r-MtL. This appears to have yielded a purified r-MtL preparation with a specific activity higher than that of native MtL which was isolated from cultures not supplemented with additional copper.

The ascomycete fungus *M. thermophila* produces a constellation of thermostable cellulases (48, 53, 58–60) and at least one thermotolerant xylanase (76). Whether this organism might be a good source of thermostable laccase was a subject of this investigation. The observation that r-MtL retains virtually 100% activity after 20-min incubation at 60°C seems to validate our approach. In addition, Xu et al. (71) disclose that MtL not only is more thermostable than laccases from the basidiomycetes *T. villosa* and *R. solani*, but also demonstrates a pronounced thermal activation such that preincubation at elevated temperatures gives higher activity.

The yield of r-MtL from *A. oryzae* cotransformants grown in shake flasks was modest (11 to 19 mg per liter). However, these yields are consistent with those obtained by heterologous expression of basidiomycete laccases in other hosts (35, 55). In addition, it seems likely that the heterologous expression of laccases in *A. oryzae* will benefit from the successful history of industrial scale-up, strain development, and process methods for other *Aspergillus* enzyme products.

ACKNOWLEDGMENTS

We thank Ejner B. Jensen (Novo Nordisk A/S), Alan V. Klotz, Glenn E. Nedwin, Jeffrey Shuster, Jill A. Wahleithner, and Debbie S. Yaver (Novo Nordisk Biotech, Inc.) for critical reading of the manuscript and for helpful suggestions throughout the course of this work. We are grateful to W. Shin and E. I. Solomon of Stanford University for copper atomic absorption spectroscopy of MtL, J. M. Bailey of Hewlett-Packard for C-terminal sequencing, and S. Osborn of Zymogenetics, Inc., for help with carbohydrate analysis. We also thank Howard Brody for the *A. oryzae* host strain, Debbie S. Yaver for the *N. crassa lcc-1* gene fragment, Michael Rey for pMWR3, and Suzie Otani for pSO2.

REFERENCES

1. **Agematu, H., K. Kominato, N. Shibamoto, T. Yoshioka, H. Nishida, R. Okamoto, T. Shin, and S. Murao.** 1993. Transformation of 7-(4-hydroxyphenylacetamido) cephalosporanic acid into a new cephalosporin antibiotic, 7-[1-oxaspiro(2.5)octa-6-oxo-4,7-diene-2-carboxamido]cephalosporanic acid, by laccase. Biosci. Biotech. Biochem. **57:**1387–1388.
2. **Ander, P., and K.-E. Eriksson.** 1976. The importance of phenol oxidase activity in lignin degradation by the white-rot fungus *Sporotrichum pulverulentum*. Arch. Microbiol. **109:**1–8.
3. **Apinis, A. E.** 1963. Occurrence of thermophilous microfungi in certain alluvial soils near Nottingham. Nova Hedwigia **5:**57–78.
4. **Aramayo, R., and W. E. Timberlake.** 1990. Sequence and molecular structure of the *Aspergillus nidulans yA* (lacasse I) gene. Nucleic Acids Res. **18:**3415.
5. **Bar-Nunn, N., A. Tal-Lev, E. Harel, and A. M. Mayer.** 1988. Repression of laccase formation in *Botrytis cinerea* and its possible relation to phytopathogenicity. Phytochemistry **27:**2505–2509.
6. **Bauer, R., and C. O. Rupe.** 1971. Use of syringaldazine in a photometric method for estimating "free" chlorine in water. Anal. Chem. **43:**421–425.

6a. **Berka, R.** Unpublished results.

7. **Berka, R. M., and C. C. Barnett.** 1989. The development of gene expression systems for filamentous fungi. Biotechnol. Adv. **7:**127–154.
8. **Blanchette, R. A.** 1991. Delignification by wood-decay fungi. Annu. Rev. Phytopathol. **29:**381–398.
9. **Bligny, R., and R. Douce.** 1983. Excretion of laccase by sycamore (*Acer pseudoplantanus*) cells. Purification and properties of the enzyme. Biochem. J. **209:**489–496.
10. **Bligny, R., J. Faillard, and R. Douce.** 1986. Excretion of laccase by sycamore (*Acer pseudoplantanus* L.) cells. Effects of a copper deficiency. Biochem. J. **237:**583–588.
11. **Boel, E., T. Christensen, and H. F. Wöldlike.** 1996. Process for the production of protein products in *Aspergillus oryzae* and a promoter for use in *Aspergillus*. U.S. Patent 5536661.
12. **Bolivar, F., R. L. Rodriguez, P. J. Greene, M. C. Betlach, H. L. Heyneker, H. W. Boyer, J. H. Crosa, and S. Falkow.** 1977. Construction and characterization of new cloning vehicles. II. A multipurpose cloning system. Gene **2:**95–113.
13. **Bollag, J.-M.** 1992. Decontaminating soil with enzymes. Environ. Sci. Technol. **26:**1876–1881.
14. **Bollag, J.-M., and A. Leonowicz.** 1984. Comparative studies of extracellular fungal laccases. Appl. Environ. Microbiol. **48:**849–854.
15. **Bollag, J.-M., and C. Myers.** 1992. Detoxification of aquatic and terrestrial sites through binding of pollutants to humic substances. Sci. Total Environ. **117/118:**357–366.
16. **Bollag, J.-M., K. L. Shuttleworth, and D. H. Anderson.** 1988. Laccase-mediated detoxification of phenolic compounds. Appl. Environ. Microbiol. **54:**3086–3091.
17. **Bourbonnais, R., and M. G. Paice.** 1992. Oxidation of non-phenolic substrates. An expanded role for laccase in lignin biodegradation. FEBS Lett. **267:**99–102.
18. **Childs, R. E., and W. G. Bardsley.** 1975. The steady-state kinetics of peroxidase with 2,2′-azino-di-(3-ethylbenzthiazoline-6-sulfonic acid) as a chromagen. Biochem. J. **145:**93–103.
19. **Choi, G. H., T. G. Larson, and D. L. Nuss.** 1992. Molecular analysis of the laccase gene from the chestnut blight fungus and selective suppression of its expression in an isogenic hypovirulent strain. Mol. Plant-Microbe Interact. **5:**119–128.
20. **Christensen, T.** 1994. Application: *Aspergillus oryzae* as a host for production of industrial enzymes. FEMS Symp. **69:**251–259.
21. **Christensen, T., H. Wöldlike, E. Boel, S. B. Mortensen, K. Hjortshøj, L. Thim, and M. T. Hansen.** 1988. High level expression of recombinant genes in *Aspergillus oryzae*. Bio/Technology **6:**1419–1422.
22. **Clutterbuck, A. J.** 1972. Absence of laccase from yellow-spored mutants of *Aspergillus nidulans*. J. Gen. Microbiol. **70:**423–425.
23. **Coll, P. M., C. Tabernero, R. Santamaría, and P. Pérez.** 1993. Characterization and structural analysis of the laccase I gene from the newly isolated ligninolytic basidiomycete PM1 (CECT 2971). Appl. Environ. Microbiol. **59:**4129–4135.
24. **Davies, R. W.** 1991. Molecular biology of a high-level recombinant protein production system in *Aspergillus*, p. 45–81. *In* S. A. Leong and R. M. Berka (ed.), Molecular industrial mycology. Systems and applications for filamentous fungi. Marcel Dekker, New York, N.Y.
25. **Davis, R. W., D. Botstein, and J. R. Roth.** 1980. Advanced bacterial genetics: a manual for genetic engineering. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
26. **Dec, J., and J.-M. Bollag.** 1990. Detoxification of substituted phenols by oxidoreductive enzymes through polymerization reactions. Arch. Environ. Contam. Toxicol. **19:**543–550.
27. **Felsenfeld, G.** 1960. The determination of cuprous ion in copper proteins. Arch. Biochem. Biophys. **87:**247–251.
28. **Fernandez-Larrea, J., and U. Stahl.** 1996. Isolation and characterization of a laccase gene from *Podospora anserina*. Mol. Gen. Genet. **252:**539–551.
29. **Germann, U. A., G. Müller, P. E. Hunziker, and K. Lerch.** 1988. Characterization of two allelic forms of *Neurospora crassa* laccase. J. Biol. Chem. **263:**885–896.
30. **Giesecke, H., B. Obermaier, H. Domedy, and W. J. Neubert.** 1992. Rapid sequencing of the Sendai virus 6.8 kb large L gene through primer walking with an automated DNA sequencer. J. Virol. Methods **38:**47–60.
31. **Gurr, S. J., S. E. Unkles, and J. R. Kinghorn.** 1987. The structure and organization of nuclear genes in filamentous fungi, p. 93–139. *In* J. R. Kinghorn (ed.), Gene structure in eukaryotic microbes. IRL Press, Oxford, United Kingdom.

32. **Iimura, Y., K. Takaneouchi, M. Nakamura, S. Kawai, Y. Katayama, and N. Morohoshi.** 1992. Cloning and sequence analysis of laccase genes and its use for an expression vector in *Coriolus versicolor*, p. 427–431. *In* M. Kuwahara and M. Shimada (ed.), Proceedings of the Fifth International Conference on Biotechnology in the Pulp and Paper Industry. Uni Publishers Co. Ltd., Tokyo, Japan.
33. **Jönsson, L., K. Sjöström, I. Häggström, and P. O. Nyman.** 1995. Characterization of a laccase gene from the white-rot fungus *Trametes versicolor* and structural features of basidiomycete laccases. Biochim. Biophys. Acta **1251:** 210–215.
34. **Kantelinen, A., A. Hataka, and L. Vilkari.** 1989. Production of lignin peroxidase and laccase by *Phlebia radiata*. Appl. Microbiol. Biotechnol. **31:**234–239.
35. **Kojima, Y., Y. Tsukuda, Y. Kawai, A. Tsukamoto, J. Sugiura, M. Sakaino, and Y. Kita.** 1990. Cloning, sequence analysis, and expression of ligninolytic polyphenoloxidase genes of the white-rot basidiomycete *Coriolus hirsutus*. J. Biol. Chem. **265:**15224–15230.
36. **Korman, D. R., F. T. Bayliss, C. C. Barnett, C. L. Carmona, K. H. Kodama, T. J. Royer, S. A. Thompson, M. Ward, L. J. Wilson, and R. M. Berka.** 1990. Cloning, characterization and expression of two α-amylase genes from *Aspergillus niger* var. *awamori*. Curr. Genet. **17:**203–212.
37. **Lante, A., A. Crapisi, G. Pasini, A. Zamorani, and P. Spettoli.** 1992. Immobilized laccase for must and wine processing, p. 558–562. *In* D. S. Clark and D. A. Estell (ed.), Enzyme engineering XI. The New York Academy of Sciences, New York, N.Y.
38. **Leatham, G. F., and M. A. Stahmann.** 1981. Studies on the laccase of *Lentinus edodes*: specificity, localization and association with the development of fruiting bodies. J. Gen. Microbiol. **125:**147–157.
39. **Leonowicz, A., K. Grzywnowicz, and M. Malinowska.** 1979. Oxidative and demethylating activity of multiple forms of laccase (EC 1.14.18.1) from *Pholiota mutabilis*. Acta Biochim. Polon. **26:**431–434.
40. **Lewin, B.** 1985. Genes, 2nd ed. John Wiley & Sons, New York, N.Y.
41. **Malkin, R., and B. G. Malmström.** 1970. The state and function of copper in biological systems. Adv. Enzymol. Relat. Subj. Biochem. **33:**177–244.
42. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Press, Cold Spring Harbor, N.Y.
43. **Messerschmidt, A., and R. Huber.** 1990. The blue oxidases, ascorbate oxidase, laccase, and ceruloplasmin. Modeling and structural relationships. Eur. J. Biochem. **187:**341–352.
44. **Messerschmidt, A., R. Rossi, R. Landenstein, R. Huber, M. Bolognesi, G. Gatti, A. Marchesini, R. Petruzzelli, and A. Finazzi-Agrò.** 1989. X-ray crystal structure of the blue oxidase ascorbate oxidase from *Zucchini*. Analysis of the polypeptide fold and a model of the copper sites and ligands. J. Mol. Biol. **206:**513–530.
45. **Messing, J., R. Crea, and P. H. Seeburg.** 1981. A system for shotgun DNA sequencing. Nucleic Acids Res. **9:**309–321.
46. **Nanniperri, P., and J.-M. Bollag.** 1991. Use of enzymes to detoxify pesticide-contaminated soils and waters. J. Environ. Qual. **20:**510–517.
47. **Oakley, B. R., J. E. Rinehart, B. L. Mitchell, E. Oakley, C. Carmona, G. L. Gray, and G. S. May.** 1987. Cloning, mapping and molecular analysis of the *pyrG* (orotidine-5′-phosphate decarboxylase) gene of *Aspergillus nidulans*. Gene **61:**385–399.
48. **Oberson, J., T. Binz, D. Fracheboud, and G. Canevascini.** 1992. Comparative investigation of cellulose-degrading enzyme systems produced by different strains of *Myceliophthora thermophila* (Apinis) v. Oorschot. Enzyme Microb. Technol. **14:**303–312.
49. **Perry, C. R., M. Smith, C. H. Britnell, D. A. Wood, and C. F. Thurston.** 1993. Identification of two laccase genes in the cultivated mushroom *Agaricus bisporus*. J. Gen. Microbiol. **139:**1209–1218.
50. **Reid, I. D.** 1991. Biological pulping in paper manufacture. Trends Biotechnol. **9:**262–265.
51. **Reinhammer, B.** 1970. Purification and properties of laccase and stella cyanin from *Rhus vernicifera* D. Biochim. Biophys. Acta **205:**35–47.
52. **Rowlands, T. R., and G. Turner.** 1973. Nuclear and extranuclear inheritance of oligomycin resistance in *Aspergillus nidulans*. Mol. Gen. Genet. **126:**201–216.
53. **Roy, S. K., S. K. Raha, S. K. Dey, and S. L. Chakrabarty.** 1990. Effect of temperature on the production and location of cellulase components in *Myceliophthora thermophila* D-14 (ATCC 48104). Enzyme Microb. Technol. **12:**710–713.
54. **Ruggiero, P., J. M. Sarkar, and J.-M. Bollag.** 1989. Detoxification of 2,4-dichlorophenol by a laccase immobilized on soil or clay. Soil Sci. **147:**361–370.
55. **Saloheimo, M., and M.-L. Niku-Paavola.** 1991. Heterologous production of a ligninolytic enzyme: expression of the *Phlebia radiata* laccase gene in *Trichoderma reesei*. Bio/Technology **9:**987–990.
56. **Saloheimo, M., M.-L. Niku-Paavola, and J. K. C. Knowles.** 1991. Isolation and structural analysis of the laccase gene from the lignin-degrading fungus *Phlebia radiata*. J. Gen. Microbiol. **137:**1537–1544.
57. **Sannia, G., P. Giardina, M. Luna, M. Rossi, and V. Buonocore.** 1986. Laccase from *Pleurotus ostreatus*. Biotechnol. Lett. **8:**797–800.
58. **Sen, S., T. K. Abraham, and S. L. Chakrabarty.** 1981. Utilization of cellulosic wastes by thermophilic fungi. Adv. Biotechnol. **2:**633–637.
59. **Sen, S., T. K. Abraham, and S. L. Chakrabarty.** 1982. Characteristics of the cellulase produced by *Myceliophthora thermophila* D-14. Can. J. Microbiol. **28:**271–277.
60. **Sen, S., T. K. Abraham, and S. L. Chakrabarty.** 1983. Induction of cellulase in *Myceliophthora thermophila* D-14. Can. J. Microbiol. **29:**1258–1260.
61. **Simmons, K. E., R. D. Minard, and J.-M. Bollag.** 1989. Oxidative co-oligomerization of guiacol and 4-chloroaniline. Environ. Sci. Technol. **23:**115–121.
62. **Sorge, J. A.** 1988. Bacteriophage lambda cloning vectors, p. 43–60. *In* R. L. Rodriguez and D. T. Denhardt (ed.), Vectors, a survey of molecular cloning vectors and their uses. Butterworths, Boston, Mass.
63. **Sugiura, J., M. Sakaino, Y. Kojima, K. Tsujioka, Y. Mutou, Y. Shinohara, and K. Koide.** 1987. Purification and properties of phenol oxidases produced by white rot fungi and molecular cloning of phenol oxidase genes, p. 317–320. *In* International Seminar on Lignin Enzymic and Microbial Degradation. International Symposium on Wood and Pulping Chemistry, Paris. INRA Publications, Versailles, France.
64. **Timberlake, W. E., and E. C. Barnard.** 1981. Organization of a gene cluster expressed specifically in the asexual spores of *Aspergillus nidulans*. Cell **26:** 29–37.
65. **van Oorschot, C. A. N.** 1977. The genus *Myceliophthora*. Persoonia **9:**401–408.
66. **Vieira, J., and J. Messing.** 1987. Production of single-stranded plasmid DNA. Methods Enzymol. **153:**3–11.
67. **von Heijne, G.** 1984. How signal sequences maintain cleavage specificity. J. Mol. Biol. **173:**243–251.
68. **von Klopotek, A.** 1974. Revision of thermophilic *Sporotrichum*-spp.: *Chrysosporium thermophilum* (Apinis) comb. nov. and *Chrysosporium fergusil* sp. nov. Conidial state of *Corynascus thermophilus* comb. nov. Arch. Microbiol. **98:**365–369.
69. **Wahleithner, J. A., F. Xu, K. M. Brown, S. H. Brown, E. J. Golightly, T. Halkier, S. Kauppinen, A. Pederson, and P. Schneider.** 1996. The identification and characterization of four laccases from the plant pathogenic fungus *Rhizoctonia solani*. Curr. Genet. **29:**395–403.
70. **Xu, F.** 1997. Effects of redox potential and hydroxide inhibition on the pH activity profile of fungal laccases. J. Biol. Chem. **272:**924–928.
71. **Xu, F., W. Shin, S. H. Brown, J. A. Wahleithner, U. M. Sundaram, and E. I. Solomon.** 1996. A study of a series of fungal laccases and bilirubin oxidase that exhibit significant differences in redox potential, substrate specificity, and stability. Biochim. Biophys. Acta **1292:**303–311.
72. **Yanisch-Perron, C., J. Vieira, and J. Messing.** 1985. Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. Gene **33:**103–119.
73. **Yaropolov, A. I., O. V. Skorobogatko, S. S. Vartonov, and S. D. Varfolomeyev.** 1994. Laccase. Properties, catalytic mechanism, and applicability. Appl. Biochem. Biotechnol. **49:**257–280.
74. **Yasuchi, K., K. Yukio, and T. Yukiko.** September 1990. DNA for expression and secretion. European patent application EP 0 388 166 A1.
75. **Yaver, D. S., F. Xu, E. J. Golightly, K. M. Brown, S. H. Brown, M. W. Rey, P. Schneider, T. Halkier, K. Mondorf, and H. Dalbøge.** 1996. Purification, characterization, molecular cloning, and expression of two laccase genes from the white-rot basidiomycete *Trametes villosa*. Appl. Environ. Microbiol. **62:**834–841.
76. **Zamost, B. L., H. K. Nielsen, and R. L. Starnes.** 1991. Thermostable enzymes for industrial applications. J. Ind. Microbiol. **8:**71–82.

APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Nov. 1998, p. 4423–4427
0099-2240/98/$04.00+0

Vol. 64, No. 11

Molecular Characterization and Expression of a Phytase Gene from the Thermophilic Fungus *Thermomyces lanuginosus*

RANDY M. BERKA, MICHAEL W. REY, KIMBERLY M. BROWN, TONY BYUN, AND ALAN V. KLOTZ*
Novo Nordisk Biotech, Davis, California 95616-4880

Received 17 February 1998/Accepted 10 September 1998

The *phyA* gene encoding an extracellular phytase from the thermophilic fungus *Thermomyces lanuginosus* was cloned and heterologously expressed, and the recombinant gene product was biochemically characterized. The *phyA* gene encodes a primary translation product (PhyA) of 475 amino acids (aa) which includes a putative signal peptide (23 aa) and propeptide (10 aa). The deduced amino acid sequence of PhyA has limited sequence identity (ca. 47%) with *Aspergillus niger* phytase. The *phyA* gene was inserted into an expression vector under transcriptional control of the *Fusarium oxysporum* trypsin gene promoter and used to transform a *Fusarium venenatum* recipient strain. The secreted recombinant phytase protein was enzymatically active between pHs 3 and 7.5, with a specific activity of 110 μmol of inorganic phosphate released per min per mg of protein at pH 6 and 37°C. The *Thermomyces* phytase retained activity at assay temperatures up to 75°C and demonstrated superior catalytic efficiency to any known fungal phytase at 65°C (the temperature optimum). Comparison of this new *Thermomyces* catalyst with the well-known *Aspergillus niger* phytase reveals other favorable properties for the enzyme derived from the thermophilic gene donor, including catalytic activity over an expanded pH range.

Phytases (*myo*-inositol hexakisphosphate phosphohydrolases; EC 3.1.3.8) catalyze the hydrolysis of phytic acid (*myo*-inositol hexakisphosphate) to the mono-, di-, tri-, tetra-, and pentaphosphates of *myo*-inositol and inorganic phosphate. A broad range of microorganisms, including bacteria (20), yeasts (2), and filamentous fungi (10, 19, 27), produce phytases.

Phytic acid is the primary storage form of phosphate in cereal grains, legumes, and oilseeds, such as soy, which are the principal components of animal feeds. However, monogastric animals are unable to metabolize phytic acid and largely excrete it in their manure. Therefore, the presence of phytic acid in animal feeds for chickens and pigs is undesirable, because the phosphate moieties of phytic acid chelate essential minerals and possibly proteins, rendering the nutrients unavailable. Since phosphorus is an essential element for the growth of all organisms, livestock feed must be supplemented with inorganic phosphate. There are a number of published reports (12, 16, 18, 26) describing the use of phytases in the feeds of monogastric animals and in human food.

When phytic acid is not metabolized by monogastric animals the phosphate level in the manure can also create disposal problems. The amount of manure produced worldwide has increased significantly as a result of increased livestock production. Environmental pollution with high-phosphate manure has caused problems in various locations around the world due to the accumulation of phosphate, particularly in bodies of water. Consequently, animal feed distributors in Europe have begun to formulate feed products with supplemental phytase in order to improve feedlot productivity and decrease phosphate waste. Thus, phytases are also useful for reducing the amount of phytate in manure (13, 18). The current commercial feed supplement is a recombinant *Aspergillus niger* (previously *Aspergillus ficuum*) phytase produced in *Aspergillus niger* (27) or *Aspergillus oryzae* (i.e., Phytase Novo [13]).

* Corresponding author. Mailing address: Novo Nordisk Biotech, Inc., 1445 Drew Ave., Davis, CA 95616-4880. Phone: (530) 757-0822. Fax: (530) 758-0317. E-mail: magi@nnbt.com.

There is a definite commercial need for second-generation phytases with improved properties (e.g., higher thermostability and catalytic efficiency) that can be produced in commercially significant quantities. Our objectives were to identify, clone, and characterize a phytase from a thermophilic fungus in anticipation that this enzyme would offer superior biochemical properties.

MATERIALS AND METHODS

DNA extraction and hybridization analysis. Total cellular DNA was extracted from *Thermomyces lanuginosus* CBS 586.94 by the procedure described by Timberlake and Bernard (21). Genomic DNA samples were analyzed by Southern hybridization (6) under conditions of low stringency (i.e., 5× SSPE [1× SSPE is 0.18 M NaCl, 10 mM NaH_2PO_4, and 1 mM EDTA {pH 7.7}], 25% formamide, 0.3% sodium dodecyl sulfate [SDS]). A phytase-specific probe fragment comprising the *Aspergillus niger phyA* coding region (approximately 1.6 kb) was radiolabeled by nick translation (11) with [α-^{32}P]dCTP (Amersham, Arlington Heights, Ill.) and added to the hybridization buffer at an activity of approximately 10^6 cpm per ml. The hybridization and washing conditions have been described previously (4).

DNA libraries and identification of phytase clones. Genomic DNA libraries were constructed with the bacteriophage cloning vector λZipLox (Life Technologies, Gaithersburg, Md.) with *Escherichia coli* Y1090ZL cells (Life Technologies) as a host for plating and purification of recombinant bacteriophages and *E. coli* DH10Bzip (Life Technologies) for excision of individual pZL1-phytase clones. Total cellular DNA was partially digested with *Tsp*5091 and size fractionated on 1% agarose gels. DNA fragments migrating in the range of 3 to 7 kb were excised and eluted from the gel with Prep-a-Gene reagents (Bio-Rad Laboratories, Hercules, Calif.). The eluted DNA fragments were ligated with *Eco*RI-cleaved and dephosphorylated λZipLox vector arms (Life Technologies), and the ligation mixtures were packaged with commercial packaging extracts (Stratagene, La Jolla, Calif.). The packaged DNA libraries were plated and amplified in *E. coli* Y1090ZL cells (Life Technologies). Approximately 30,000 plaques from the library were screened by plaque hybridization with the radiolabeled phytase probe. One positive clone which hybridizes strongly to the probe was picked and purified twice in *E. coli* Y1090ZL cells. The phytase clone was subsequently excised from the λZipLox vector as a pZL1-phytase clone (5) and designated pMWR46.

Molecular analysis of the *T. lanuginosus* phytase gene. Restriction mapping of pMWR46 was performed by standard methods (11). DNA sequencing of the phytase clones was performed with model 373A automated DNA sequencer (Applied Biosystems, Inc., Foster City, Calif.) by the primer-walking technique with dye-terminator chemistry (7). In addition to the *lac* forward and *lac* reverse primers, specific oligonucleotide sequencing primers were synthesized on an

Applied Biosystems model 394 DNA-RNA synthesizer according to the manufacturer's instructions.

Construction of the phytase expression vector pMWR48. The coding region of the *T. lanuginosus phyA* gene was amplified by PCR with the forward primer 5′-ATTTAAATGGCGGGGATAGGTTTGG-3′ and the reverse primer 5′-CTTAATTAATCAAAAGCAGCGATCCC-3′. The sense primer incorporated the first in-frame ATG and extends 16 bp downstream. The antisense primer incorporated a region 14 bp upstream of the translational stop codon and extends through the stop codon. To facilitate the cloning of the amplified fragment, the sense and antisense primers contain a *Swa*I and a *Pac*I restriction site, respectively. The amplified product was digested with *Swa*I and *Pac*I and ligated with pDM181 (also digested with *Swa*I and *Pac*I), a plasmid which provides the *Fusarium oxysporum* trypsin gene promoter and terminator and the *bar* resistance cassette (3). The resulting expression vector was designated pMWR48.

Transformation of *Fusarium venenatum* and analysis of transformants. Transformation protocols and methods for purification of *F. venenatum* (28) transformants are described by Royer et al. (15). Mycelia from primary transformants were used to inoculate shake flasks containing 25 ml of M400 Da medium (50 g of maltodextrin, 2 g of $MgSO_4 \cdot 7H_2O$, 2 g of KH_2PO_4, 4 g of citric acid, 8 g of yeast extract, 2 g of urea, and 0.5 ml of trace metal solution per liter [15]) and incubated with shaking at 30°C. One milliliter of culture supernatant was harvested at 4, 5, and 7 days and stored at 4°C. Phytase activity was assayed as described below. Spores from the primary transformants producing the highest phytase activity were generated by inoculating 20 ml of R medium (12.1 g of $NaNO_3$/liter, 50 g of succinic acid/liter, 20 ml of 50× Vogel's salts, 25 mM $NaNO_3$ [pH 6.0] [15]) with mycelia and incubating it at 30°C with shaking for 2 to 3 days. Single spores were isolated by spreading 150 ml of spore culture onto manipulator plates (1X Vogel's salts, 25 ml of $NaNO_3$, 2.5% sucrose, 2% Noble agar) containing 5 mg of Basta [phosphinothricin or 2-amino-4-(hydroxymethylphosphinyl)butanoic acid; Hoechst-Schering, Rodovre, Denmark] per ml and using a micromanipulator to transfer single spores to a clear region of the plate. After 3 days of growth at room temperature, the germinated spores were transferred to individual Vogel plates containing 5 mg of Basta/ml. Shake flasks containing 25 ml of M400Da medium plus 5 mg of Basta/ml were inoculated in duplicate with mycelial plugs from each single-spore isolate and incubated at 30°C. The best single-spore isolate was selected based on assay of the secreted enzymatic activity, where the transformants produced >150-fold more phytase activity than an untransformed control.

Protein purification. The best *F. venenatum* transformant was run in two 2-liter fermentors with a standard protocol (3). The frozen cell-free broth (1,700 ml) was thawed, clarified by centrifugation, and concentrated on a hollow-fiber Amicon filtration unit with an S1Y10 filter to a volume of 350 ml. The sample was adjusted to pH 7, diluted to a conductivity of 2 mS, and chromatographed at room temperature on a 75-ml-bed-volume Q-Sepharose Big Beads column (Pharmacia), which had been equilibrated in 20 mM Tris-Cl, pH 7. The column was developed at 5 ml/min with the equilibration buffer until the effluent A_{280} had decreased to near baseline. The column was then developed at 5 ml/min with a 600-ml gradient of 0 to 0.6 M NaCl in the same buffer. The bound enzyme activity was found to elute in fractions corresponding to ca. 0.2 M NaCl.

The collected activity peak was concentrated by ultrafiltration with a PM-10 membrane to a volume of 25 ml, diluted to a conductivity of 0.9 mS, and chromatographed at 4 ml/min on a MonoQ HR 10/16 column which had been equilibrated in 20 mM MOPS (morpholinepropanesulfonic acid), pH 7. The column was developed with 80 ml of starting buffer and then with a 400-ml gradient of 0 to 0.5 M NaCl in the same buffer. Enzyme activity was detected in fractions by using the *p*-nitrophenyl phosphate measurement described below. The active fractions were also analyzed with a Novex 10 to 27% gradient SDS-polyacrylamide gel, and the fractions were combined if judged by electrophoresis to be substantially purified.

The peak fractions were combined, concentrated with an Amicon PM-10 membrane by ultrafiltration, and exchanged into 20 mM MES (morpholine ethanesulfonic acid), pH 5.5. The sample conductivity was 1.1 mS. One-third of this sample was chromatographed at 1 ml/min on a Mono S HR 5/5 column (Pharmacia) which had been equilibrated in the same buffer. The column was developed with 5 ml of starting buffer and then with a 25-ml linear gradient of 0 to 0.6 M sodium chloride in the same buffer. The active fractions were combined after electrophoretic analysis to eliminate those which contained trace contaminants.

Physicochemical characterization. Isoelectric focusing (IEF) was performed with a Novex pH 3 to 7 IEF gel according to the instructions of the manufacturer. IEF standards from both Pharmacia and Bio-Rad were used to calibrate the gel.

The protein extinction coefficient was determined experimentally by quantitative amino acid analysis with a Hewlett-Packard AminoQuant system. The analysis assumed 49,700 for the protein molecular weight, based on the translated gene sequence for the mature protein.

Amino-terminal sequence analysis was performed on an Applied Biosystems 476A sequencer.

Enzyme assays. Phytase activity was measured by two different methods. During purification, fractions were rapidly evaluated by measuring the rate of *p*-nitrophenyl phosphate hydrolysis at 405 nm with 10 mM substrate in 0.2 M sodium citrate, pH 5.5, at 30°C with a plate reader (Thermomax; Molecular Devices).

Enzyme kinetics studies performed on purified enzyme samples were accomplished by the assay of inorganic phosphate liberated from corn phytic acid (Sigma catalog no. P 8810). Exhaustive phytate hydrolysis was accomplished by incubating 0.5 or 0.1% phytic acid with enzyme (1 U/ml) in 0.2 M sodium citrate, pH 5.5, at 37°C. Aliquots were removed over a period of 10 h and analyzed (see below) for kinetics of phosphorus release. Ten hours was found to be sufficient for the completion of product formation. Standard enzyme kinetics reactions were carried out for 30 min at 37°C in 0.5% (wt/wt) phytic acid. The reaction was quenched by the addition of an equal volume of 15% (wt/wt) trichloroacetic acid. After cooling, 100 μl of the resulting mixture was diluted in 1 ml of water. The sample was incubated at 50°C for 5 min. Color reagent (1 ml) was added, and the 50°C incubation was continued for 15 min. The absorbance of a 200-μl aliquot was measured at 690 nm with a microplate reader. The color reagent was composed of 6 N sulfuric acid–water–2.5% (wt/vol) hepta-ammonium molybdate–10% ascorbate (aqueous) in a ratio of 1:2:1:1 and was prepared fresh daily. Quantitation was based on a standard curve generated with a 10 mM sodium monobasic phosphate standard. One unit is defined as 1 μmol of inorganic phosphate released per min with 0.5% phytic acid in 0.2 M sodium citrate, pH 5.5, at 37°C.

Steady-state kinetics measurements were made by substrate titration. Phytate concentrations were 2.16, 1.08, 0.541, 0.216, 0.108, and 0.0758 mM for K_m determination. Phytate concentrations of 1.08, 0.541, 0.216, and 0.108 mM in the presence or absence of 1 mM sodium monobasic phosphate were used to evaluate product inhibition.

Thermostability measurement. Phytase samples were dissolved at 100 U per ml in 0.2 M sodium citrate, pH 5.5. One hundred-microliter aliquots of each enzyme solution were incubated for 20 min in a water bath at 37, 45, 50, 55, 60, 65, 70, and 75°C. After the heat treatment, the samples were stored at 0°C until activity assays were performed. Each sample was diluted 1:80 in 0.2 M sodium citrate, pH 5.5, containing 0.01% (wt/wt) Tween 20, and the standard activity assay was performed.

pH-activity measurement. To attain a buffering range between pHs 2 and 7, a three-component 125 mM glycine-acetate-citrate buffer was employed. The buffer components were combined at final concentrations of 42 mM per component, and phytic acid was added as a solid to 1% (wt/wt). This mixture was adjusted to pH 7 with concentrated HCl, and a 10-ml aliquot was taken. This process was repeated for every 0.5 pH units through pH 2.

Enzyme stock solutions of 20 U per ml were prepared in 20 mM MES buffer, pH 5.5. Substrate (1% [wt/wt]; 850 μl) in buffer at a given pH was combined with 100 μl of water and 50 μl of enzyme stock solution and incubated for 30 min at 37°C. Subsequently, the enzyme reaction was quenched with 1 ml of 15% trichloroacetic acid and quantitated by the standard method.

Temperature-activity measurement. Enzyme stock solutions of 12.5 U per ml were prepared in 0.2 M sodium citrate buffer, pH 5.5. Two hundred fifty microliters of 1% phytic acid substrate was added to a 1.7-ml Eppendorf tube followed by 240 μl of 0.2 M sodium citrate buffer, pH 5.5. This solution was vortexed and placed in a water bath at the designated temperature. After 20 min of equilibration in the water bath, the mixture was vortexed and 10 μl of phytase solution was added. The sample was vortexed and incubated in the water bath for an additional 30 min, and then the reaction was quenched with 1 ml of 15% trichloroacetic acid and quantitated by the standard method.

Nucleotide sequence accession number. The complete *phyA* gene sequence has been deposited in GENESEQN as accession no. T90070.

RESULTS

Cloning of phytase gene sequences from *T. lanuginosus*. Southern blotting experiments indicated that an *Aspergillus* phytase gene fragment could be used as a probe to identify phytase gene-specific fragments in *T. lanuginosus* genomic DNA (Fig. 1). We screened 30,000 plaques from a genomic library of *T. lanuginosus* DNA constructed in λZipLox for hybridization with the *Aspergillus* phytase gene probe. Several positive clones were picked and excised by an in vivo-excision protocol (5).

Analysis of the *T. lanuginosus phyA* gene. DNA sequencing of one *T. lanuginosus* phytase clone (pMWR46) showed an open reading frame similar to the *A. niger* phytase gene. The positions of introns and exons within the *phyA* gene were assigned based on comparison of the deduced amino acid sequence with the deduced amino acid sequence of the corresponding *A. niger* phytase gene product. On the basis of this analysis, the *T. lanuginosus* phytase gene is comprised of two exons (47 and 1,377 bp), which are separated by a small intron (56 bp). The size and composition of the intron is consistent with those of other fungal genes (9) in that all contain consen-

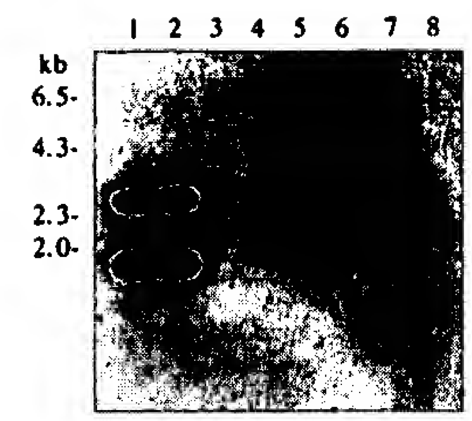


FIG. 1. Autoradiogram from Southern hybridization analysis of *T. lanuginosus* genomic DNA with an *Aspergillus* phytase gene probe. Lanes 1 and 2, *A. niger* genomic DNA digested with *Bam*HI and *Bam*HI plus *Pst*I, respectively; lanes 3 and 4, *Myceliophthora thermophila* genomic DNA digested with *Bam*HI and *Bam*HI plus *Pst*I, respectively; lanes 5 and 6, *Thielavia terrestris* genomic DNA cleaved with *Bam*HI and *Bam*HI plus *Pst*I, respectively; lanes 7 and 8, *T. lanuginosus* genomic DNA cut with *Bam*HI and *Bam*HI plus *Pst*I, respectively.

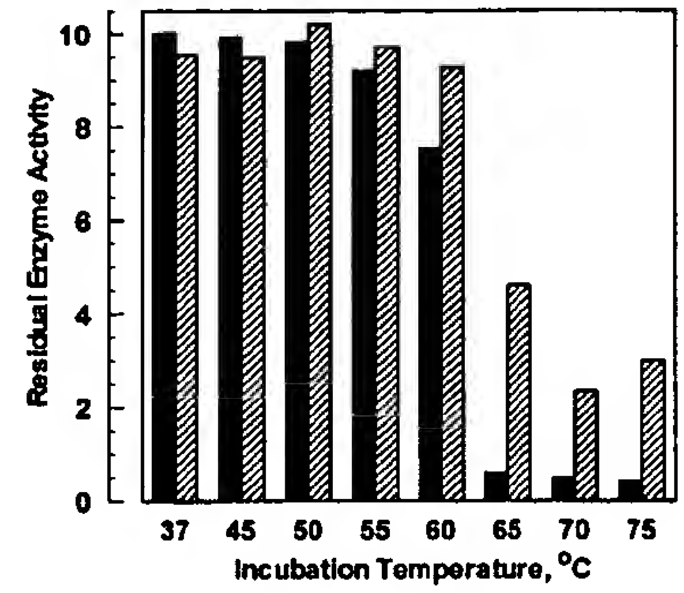


FIG. 3. Phytase thermal stability. Comparison of residual enzyme activity after a 20-min incubation at various temperatures. Full activity corresponds to 10 U. Solid bar, *A. niger* phytase; cross-hatched bar, *T. lanuginosus* phytase.

sus splice donor and acceptor sequences as well as a near approximation of the consensus lariat sequence (RCTRAC) near the 3′ end of each intervening sequence.

The deduced amino acid sequence of the *T. lanuginosus* gene product shows the characteristics of an extracellular fungal enzyme with a cleavable signal sequence. Based on the rules of von Heijne (25), the first 22 amino acids of PhyA likely comprise a secretory signal peptide which directs the nascent polypeptide into the endoplasmic reticulum. Amino-terminal amino acid sequencing suggests that the next 10 amino acids constitute a propeptide which terminates with a dibasic cleavage site (LysLys). The mature PhyA is an acidic protein (predicted isoelectric point, 5.4) composed of 452 amino acids (molecular mass, 51 kDa). The amino acid sequence also contains the active-site motif RHGXRXP, which is shared by other known phytases and acid phosphatases (Fig. 2) (23, 27). Lastly, the deduced amino acid sequence of the mature PhyA has approximately 47.5% identity with the phytase from *A. niger* (GenBank accession no. M94550).

Analysis of *F. venenatum* transformants expressing *T. lanuginosus* phytase. *F. venenatum* has recently been developed as an efficient fungal host for the production of heterologous proteins (15). Culture supernatants from 14 of the 17 primary transformants of pMWR48 were positive when assayed for phytase activity. Two primary transformants with the highest phytase activity were selected for single-spore isolation, and nine single-spore isolates were obtained.

Physicochemical characterization of the recombinant phytase. The purified *T. lanuginosus* phytase was apparently homogeneous in SDS-polyacrylamide gel electrophoresis, with a single component corresponding to a molecular weight of 60,000. The protein sample contained numerous components in IEF analysis ranging from pH 4.7 to 5.2. In contrast to the *T. lanuginosus* phytase, recombinant *A. niger* phytase is composed of a single major component with a pI near 4.9 and two minor bands around pI 4.7.

```
Thermomyc PhyA    (33)  C R V E F V Q V L S R H G A R Y P T A H K S E
Myceliophth PhyA  (45)  C E V T F A Q V L S R H G A R A P T L K R A A
Talaromyc PhyA    (54)  C K I T F V Q L L S R H G A R Y P T S S K T E
A.fumigatus PhyA  (44)  C R I T L V Q V L S R H G A R Y P T S S K S K
A.ficuum PhyA     (52)  C R V T F A Q V L S R H G A R Y P T D S K G K
A.ficuum AP2.5    (52)  C E V D T V I M V K R H G E R Y P S P S A G K
YScAP3            (46)  C E M K Q L Q M L A R H G E R Y P T Y S K G A
YScAP5            (46)  C E M K Q L Q M L A R H G E R Y P T V S L A K
HuPAP             (1)   K E L K F V T L V F R H G D R S P I D T F P T
HuLAP             (1)   R S L R P V T L L Y R H G D R S P V K T I P K
E.coliAP          (6)   L K L E S V V I V S R H G V R A P T K A T Q L
CONSENSUS                                   R H G X R X P
```

FIG. 2. Alignment of putative active-site regions of acid phosphatases (AP) and phytases from various species. The *M. thermophila* (*Myceliophth*; TREMBL O00107), *Talaromyces thermophilus* (*Talaromyc*; TREMBL O00096), *A. fumigatus* (TREMBL O00092), *A. ficuum* (*A. niger*) (SwissProt P34752 and P34754), *Saccharomyces cerevisiae* (YScAP3 and -5; SwissProt P24031 and P00635), human (HuPAP and HuLAP; SwissProt P15309 and P11117), and *E. coli* (SwissProt P07102) sequences were obtained from the databases indicated. The numbers in parentheses are the starting amino acid positions from the mature proteins for the sequences compared. Identical amino acids are boxed. *Thermocyl*, *T. lanuginosus*.

Amino-terminal sequence analysis of the purified *T. lanuginosus* enzyme identified three components: the major component (ca. 60%) is H_2N-His-Pro-Asn-Val-Asp-Ile-Ala-Arg-His-Trp-Gly-Gln. . ., which corresponds to a Kex2 cleavage site at position 34 in the primary translation product. Two minor sequences, H_2N-Gly-Glu-Asp-Glu-Pro-Phe-Val-Arg-Val-Leu-Val-Asn. . .(ca. 30%) and H_2N-Ser-Glu-Glu-Glu-Glu-Glu-Gly-Glu-Asp-Glu-Pro-Phe. . .(ca. 10%), correspond to internal cleavage sites near the COOH terminal of the protein at positions 428 and 435 in the primary translation product. The observation that our protein sequence data exactly match the predicted translation product of the *T. lanuginosus* gene and the finding that untransformed *Fusarium* host strains produce 2 orders of magnitude less enzyme activity both argue strongly that we have isolated a heterologous gene product.

The specific activities for the two recombinant phytases (i.e., those of *T. lanuginosus* and *A. niger*) were 91 and 180 U/mg, respectively, under standard assay conditions at pH 5.5. At its pH 6 optimum *T. lanuginosus* phytase had a specific activity of 110 μmol of inorganic phosphate released per min per mg of protein at 37°C. Exhaustive enzymatic hydrolysis of phytic acid revealed that *A. niger* and *T. lanuginosus* phytases released identical amounts (70%) of the total theoretically available phosphorus. Steady-state kinetic measurements disclosed that the apparent K_m of *T. lanuginosus* phytase is approximately 110 μM with respect to phytate while *A. niger* has an apparent K_m of 200 μM. There was a faint indication of excess substrate inhibition at the 2.16 mM substrate concentration, perhaps congruent with the report of inhibition above 2 mM for *A. niger* phytase (22). Steady-state kinetics measurements with 1 mM phosphate present failed to reveal any type of inhibition with this product. We estimate that the K_i for phosphate must exceed 3 mM to be undetectable in our experiments. In contrast Ullah (22) has reported that phosphate is a competitive inhibitor, with a K_i of 1.9 mM.

A comparison of enzyme thermostability profiles (Fig. 3) suggests that differences between the stabilities of the two enzymes are small. Neither enzyme is fully inactivated by a high-temperature incubation, and the residual activity profiles are consistent with partially reversible thermal denaturation

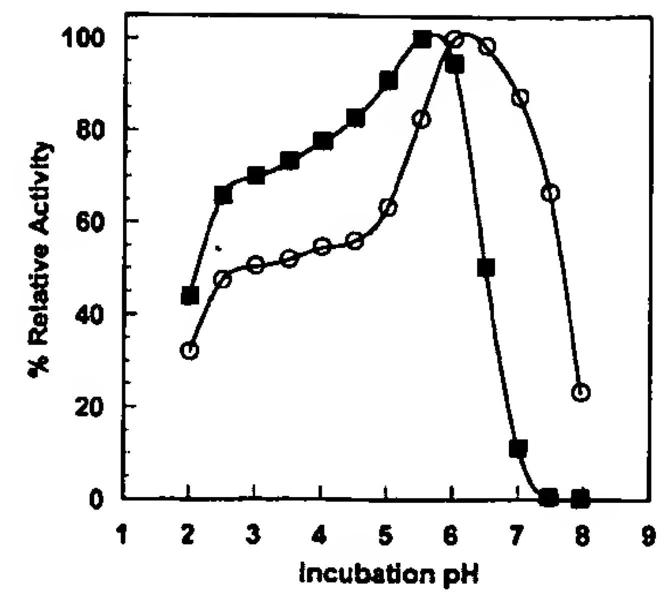


FIG. 4. Phytase pH-activity profile; comparison of relative enzyme activity at various incubation pHs. A relative activity of 100% corresponds to 1 and 1.21 μmol of inorganic phosphate released per min for *A. niger* and *T. lanuginosus* phytases, respectively. Solid square, *A. niger* phytase; open circles, *T. lanuginosus* phytase.

(24). Differential scanning calorimetry (DSC) experiments reveal that the *A. niger* enzyme has a transition at 60°C while *T. lanuginosus* phytase unfolds at 69°C. Others have reported an *Aspergillus fumigatus* phytase which has an apparently greater propensity for reversible thermal denaturation (14), as measured by residual enzyme activity. However, there are no published data on thermal denaturation points for the *A. fumigatus* phytase or other phytase species.

The pH-activity profile comparison of *T. lanuginosus* and *A. niger* phytases indicates substantial similarity between the pH profiles of the two enzymes (Fig. 4). However, the *T. lanuginosus* enzyme is active at neutral pH while the *A. niger* enzyme is not. We could not reproduce the earlier reports (e.g., reference 17) that *A. niger* phytase possesses two pH optima; employing a composite buffer, we measured a broad shoulder near pH 3. We note that there are very few cases of a single enzyme species possessing two pH optima. The earlier reports may originate from impure material which contains traces of the *A. niger* acid phosphatase (29), or they could be artifacts of employing more than one buffer to span the pH range.

Measurement of enzyme activity as a function of temperature revealed a significant difference between the two enzymes (Fig. 5). *T. lanuginosus* phytase has maximum enzyme activity near 65°C and has partial activity even at 75°C. In contrast, *A. niger* phytase is essentially inactive at 65°C. These results are congruent with the DSC data for the two enzymes, which also indicate a 9°C stability improvement for the *Thermomyces* phytase.

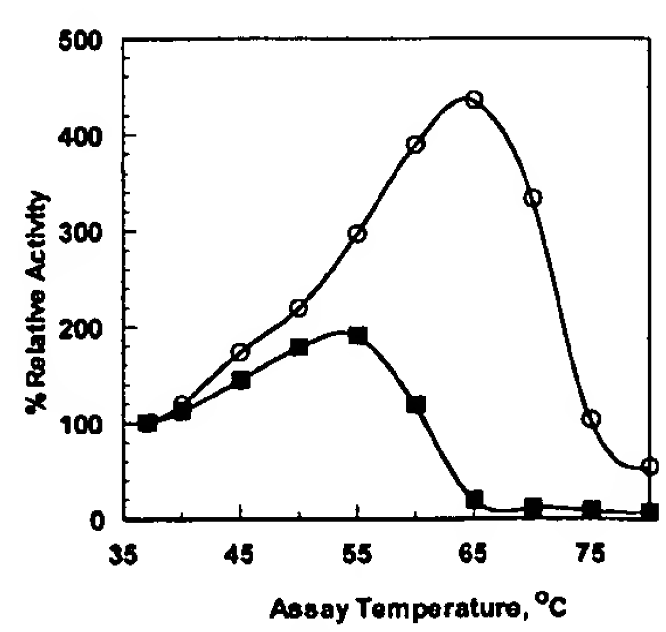


FIG. 5. Phytase temperature-activity measurement; observed enzyme activity as a function of incubation temperature. A relative activity of 100% corresponds to 0.125 μmol of inorganic phosphate released per min. Solid squares, *A. niger* phytase; open circles, *T. lanuginosus* phytase.

DISCUSSION

Enzyme activity at elevated temperatures may be relevant in applications such as saccharification (a high-temperature industrial process to generate high-fructose corn syrup), where others have reported that the addition of phytase improves carbohydrate yields (1). Figure 5 demonstrates that at 55°C, the optimal temperature for *A. niger* phytase, the *Thermomyces* phytase performs at 79% of the *A. niger* phytase turnover number (despite lower specific activity for *Thermomyces* phytase at 37°C) and at 60°C the *Thermomyces* phytase is operating at 67%-greater catalytic efficiency than the *A. niger* enzyme. The *A. niger* phytase is inactivated at 65°C, where *Thermomyces* phytase activity is maximal.

Enzyme thermal stability is also relevant in animal feed applications, where the enzyme is normally incorporated into the grains prior to pelletization and the feed briefly reaches processing temperatures of 85 to 90°C. In this circumstance a commercial phytase product must be able to withstand brief heating prior to encountering an animal's digestive tract at 37°C. Our physicochemical data demonstrate an improvement of approximately 9°C in denaturation temperature for *Thermomyces* phytase versus the present *A. niger* product.

Animal-feeding trials with formulated phytase supplementation would involve testing a total of 300 broilers or piglets at two enzyme dosages plus a control without enzyme addition. Typically the apparent total-tract digestibility of dissolved matter, organic matter, nitrogen, calcium, and total phosphorus would be monitored at one or two points during an animal's growth to determine the effect of enzyme dosage on feed intake and conversion. Such animal-feeding trials and the level of analysis required to present and evaluate the data are beyond the scope of this paper.

It is tempting to speculate about the structural origins of thermal stability in phytases. However, there is no obvious pattern to the sequence differences between phytases from thermophiles (represented by *Myceliophthora*, *Talaromyces*, and *Thermomyces*) and mesophiles (represented by *A. niger* and *A. fumigatus*). For example, there are no gross differences in protein structure, such as addition or deletion of secondary structure elements. Nor is there a systematic pattern to the sequence differences between the two representative enzymes; i.e., hydrophobic replacements, addition of salt bridges, addition of potential disulfide bonding sites, and deletion of asparagine or aspartate residues are not readily apparent. The most striking difference is the additional consensus N-linked glycosylation site present in the two *Aspergillus* enzymes (sequence position 231 in reference 27) but missing in the three thermophile examples. We believe that the most likely explanation which can be deduced for the sequence differences is derived from evolutionary rather than functional factors.

Recently the discovery of new industrial enzymes has focused on novel microbial sources representing extreme conditions (extremophiles). In many cases the genes encoding these interesting enzymes can be cloned without prior isolation of the catalyst or culturing of the donor microbe. However, heterologous production of the novel enzyme often results in extremely low yields of secreted product or accumulation of inactive material as inclusion bodies. Either of these outcomes is incompatible with the production economics required for commercialization. We have searched for new industrial catalysts from a constellation of thermophilic fungi that are more closely related than the extremophiles to the industrial fungal

production strains which are available. We have successfully isolated enzymes with both improved thermal stability characteristics and the potential for high-level commercial production (4).

T. lanuginosus phytase is an alternative enzyme with performance advantages over the conventional *A. niger* enzyme in the form of stable enzyme activity at elevated temperatures and superior substrate saturation kinetics at physiological pH. A second-generation commercial enzyme may also benefit from protein engineering when a three-dimensional protein structure is available, as is the case for the *A. fumigatus* enzyme (8).

ACKNOWLEDGMENTS

We thank Carin Morris (Novo Nordisk Biotech) for performing phytase enzyme activity assays, Sam Johnstone (Novo Nordisk Biotech) for performing *Fusarium* fermentations, and Claus Crone Fuglsang (Novo Nordisk A/S) for measuring phytase stability by DSC.

REFERENCES

1. **Antrim, R. L., C. Mitchinson, and L. P. Solheim.** 1996. Method for liquefying starch. PCT patent application WO 96/28567.
2. **Barbaric, S., B. Kozulic, B. Reis, and P. Mildner.** 1984. Physicochemical and kinetic properties of acid phosphatase from *Saccharomyces cerevisiae*. J. Biol. Chem. **259:**878–883.
3. **Berka, R. M., M. W. Rey, and A. V. Klotz.** 1997. Polypeptides having phytase activity and nucleic acids encoding same. PCT patent application WO 97/35017.
4. **Berka, R. M., P. Schneider, E. J. Golightly, S. H. Brown, M. Madden, K. M. Brown, T. Halkier, K. Mondorf, and F. Xu.** 1997. Characterization of the gene encoding an extracellular laccase of *Myceliophthora thermophila* and analysis of the recombinant enzyme produced in *Aspergillus oryzae*. Appl. Environ. Microbiol. **63:**3151–3157.
5. **D'Alessio, J. M., R. Bebee, J. L. Hartley, M. C. Noon, and D. Polayes.** 1992. Lambda ZipLox^{TM}: automatic subcloning of cDNA. Focus (Life Technologies, Gaithersburg, Md.) **14:**76–79.
6. **Davis, R. W., D. Botstein, and J. R. Roth.** 1980. Advanced bacterial genetics, a manual for genetic engineering. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
7. **Giesecke, H., B. Obermaier, H. Domedy, and W. J. Neubert.** 1992. Rapid sequencing of the Sendai virus 6.8-kb large L gene through primer walking with an automated DNA sequencer. J. Virol. Methods **38:**47–60.
8. **Grueninger-Leitch, F., A. D'Arcy, B. D'Arcy, and C. Chene.** 1996. Deglycosylation of proteins for crystallization using recombinant fusion protein glycosidases. Protein Sci. **5:**2617–2622.
9. **Gurr, S. J., S. E. Unkles, and J. R. Kinghorn.** 1987. The structure and organization of nuclear genes in filamentous fungi p. 93–139. *In* J. R. Kinghorn (ed.), Gene structure in eukaryotic microbes. IRL Press, Oxford, England.
10. **Howson, S. J., and R. P. Davis.** 1983. Production of phytate-hydrolyzing enzyme by some fungi. Enzyme Microb. Technol. **5:**377–382.
11. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
12. **Nelson, T. S., T. R. Shieh, R. J. Wodzinski, and J. H. Ware.** 1971. Effect of supplemental phytase on the utilization of phytate phosphorus by chicks. J. Nutr. **101:**1289–1294.
13. **Novo Nordisk A/S.** 1995. Phytase Novo cuts phosphorus in manure by 30%. BioTimes **10:**8–9.
14. **Pasamontes, L., M. Haiker, M. Wyss, M. Tessier, and A. P. G. H. van Loon.** 1997. Gene cloning, purification, and characterization of a heat-stable phytase from the fungus *Aspergillus fumigatus*. Appl. Environ. Microbiol. **63:**1696–1700.
15. **Royer, J. C., D. M. Moyer, S. G. Reiwitch, M. S. Madden, E. B. Jensen, S. H. Brown, C. C. Yonker, J. A. Johnstone, E. J. Golightly, W. T. Yoder, and J. R. Shuster.** 1995. *Fusarium graminearum* A 3/5 as a novel host for heterologous protein production. Bio/Technology **13:**1479–1483.
16. **Sandberg, A.-S., and H. Andersson.** 1988. Effect of dietary phytase on the digestion of phytate in stomach and small intestine of humans. J. Nutr. **118:**469–473.
17. **Sandberg, A.-S., L. R. Hulthen, and M. Türk.** 1996. Dietary *Aspergillus niger* phytase increases iron absorption in humans. J. Nutr. **126:**476–480.
18. **Schwarz, G., and P. P. Hoppe.** 1992. Phytase enzyme to curb pollution from pigs and poultry. Feed Magazine **1992:**22–26.
19. **Shieh, T. R., and J. H. Ware.** 1968. Survey of microorganisms for the production of extracellular phytase. Appl. Microbiol. **6:**1348–1351.
20. **Shimizu, M.** 1992. Purification and characterization of phytase from *Bacillus subtilis* (*natto*) N-77. Biosci. Biotechnol. Biochem. **56:**1266–1269.
21. **Timberlake, W. E., and E. C. Bernard.** 1981. Organization of a gene cluster expressed specifically in the asexual spores of Aspergillus nidulans. Cell **26:**29–37.
22. **Ullah, A. H. J.** 1988. Production, rapid purification and catalytic characterization of extracellular phytase from *Aspergillus ficuum*. Prep. Biochem. **18:**443–458.
23. **Ullah, A. H. J., and H. C. Dischinger, Jr.** 1993. Identification of active-site residues in *Aspergillus ficuum* extracellular pH 2.5 optimum acid phosphatase. Biochem. Biophys. Res. Commun. **192:**754–759.
24. **Ullah, A. H. J., and E. J. Mullaney.** 1996. Disulfide bonds are necessary for structure and activity in *Aspergillus ficuum* phytase. Biochem. Biophys. Res. Commun. **227:**311–317.
25. **von Heijne, G.** 1984. How signal sequences maintain cleavage specificity. J. Mol. Biol. **173:**243–251.
26. **Wang, H. L., E. W. Swain, and C. W. Hasseltine.** 1980. Phytase of moulds used in oriental food fermentation. J. Food Sci. **45:**1262–1266.
27. **Wodzinski, R. J., and A. H. J. Ullah.** 1996. Phytase, p. 263–302. *In* S. L. Neidleman and A. I. Laskin (ed.), Advances in applied microbiology, vol. 42. Academic Press, Inc., New York, N.Y.
28. **Yoder, W. T., and L. M. Christianson.** 1998. Species-specific primers resolve members of *Fusarium* section *Fusarium*: taxonomic status of the edible "Quorn" fungus reevaluated. Fungal Genet. Biol. **23:**68–80.
29. **Zyla, K.** 1993. The role of acid phosphatase activity during enzymic dephosphorylation of phytates by *Aspergillus niger* phytase. World J. Microbiol. Biotechnol. **9:**117–119.

THE JOURNAL OF BIOLOGICAL CHEMISTRY
© 2000 by The American Society for Biochemistry and Molecular Biology, Inc.

Vol. 275, No. 3, Issue of January 21, pp. 2009–2018, 2000
Printed in U.S.A.

Biochemical Analysis of Recombinant Fungal Mutanases

A NEW FAMILY OF α1,3-GLUCANASES WITH NOVEL CARBOHYDRATE-BINDING DOMAINS*

(Received for publication, March 23, 1999, and in revised form, August 30, 1999)

Claus C. Fuglsang‡§, Randy M. Berka¶, Jill A. Wahleithner¶||, Sakari Kauppinen‡, Jeffrey R. Shuster¶, Grethe Rasmussen‡, Torben Halkier‡, Henrik Dalbøge‡, and Bernard Henrissat‡‡**

From the ‡Novo Nordisk A/S, Bagsværd DK-2880, Denmark, ¶Novo Nordisk Biotech, Inc., Davis, California 95616, and ‡‡Architecture et Fonction des Macromolécules Biologiques, CNRS-IFR1, 13402 Marseille, France

Nucleotide sequence analysis shows that *Trichoderma harzianum* and *Penicillium purpurogenum* α1,3-glucanases (mutanases) have homologous primary structures (53% amino acid sequence identity), and are composed of two distinct domains: a NH_2-terminal catalytic domain and a putative COOH-terminal polysaccharide-binding domain separated by a *O*-glycosylated Pro-Ser-Thr-rich linker peptide. Each mutanase was expressed in *Aspergillus oryzae* host under the transcriptional control of a strong α-amylase gene promoter. The purified recombinant mutanases show a pH optimum in the range from pH 3.5 to 4.5 and a temperature optimum around 50–55 °C at pH 5.5. Also, they exhibit strong binding to insoluble mutan with K_D around 0.11 and 0.13 μM at pH 7 for the *P. purpurogenum* and *T. harzianum* mutanases, respectively. Partial hydrolysis showed that the COOH-terminal domain of the *T. harzianum* mutanase binds to mutan. The catalytic domains and the binding domains were assigned to a new family of glycoside hydrolases and to a new family of carbohydrate-binding domains, respectively.

Extracellular polysaccharides produced by microbial flora in the human oral cavity are believed to play an important role in the adherence and proliferation of bacterial aggregates on the surface of teeth (1). Consequently, these polysaccharides might have significance in the development of tartar, plaque, and possibly dental carries (2). Mutan is a major component of exopolysaccharides produced by tooth colonizing streptococci such as *Streptococcus mutans* (3). Mutan is composed of α1,3-glucan with some α1,6-glucan (dextran) side chains. Mutanase (α1,3-glucanase, EC 3.2.1.59) and dextranase (α1,6-glucanase, EC 3.2.1.11) enzymes could be beneficial additives to dentifrice preparations as it has been shown that these enzymes are capable of removing biofilms created by oral bacteria *in vitro* (4) and reducing plaque formation *in vivo* (5, 6). Mutanase activity from the filamentous fungus *Trichoderma harzianum* was first described by Guggenheim and Haller (7). However, only a limited number of reports are available on the characterization of fungal mutanases. Here we describe the cloning, expression, and subsequent characterization of two fungal mutanases representing a new family of fungal endoglucanases with their unique mutan-binding domains.

EXPERIMENTAL PROCEDURES

Fungal Strains—*T. harzianum* strain CBS 243.71 and *Penicillium purpurogenum* CBS 238.95 were used as the sources of genomic DNA. *Aspergillus oryzae* JaL142 and JaL125 (obtained from J. Lehmbeck, Novo Nordisk A/S), were alkaline protease-deficient strains used for heterologous expression of cloned mutanases. The *A. oryzae* strains show no detectable level of background mutanase activity in the assay described.

Purification and Charcaterization of the Wild-type Mutanase from T. harzianum—100 g of SP234 (Novo-Nordisk A/S, batch number PPM 3897) were dissolved in 1 liter 10 mM sodium acetate, pH 5.2. Contaminant proteins were removed by batch adsorption on DEAE-Sephadex, then by batch adsorption on S-Sepharose (Amersham Pharmacia Biotech). After concentration on a Filtron concentrator equipped with a 10-kDa cut-off membrane, the unbound material was applied to a S-Sepharose (Amersham Pharmacia Biotech) column (180 ml, 2.6 × 33 cm) equilibrated with 10 mM sodium acetate, pH 4.7. The mutanase was eluted with a 0–20 mM linear gradient of NaCl in the same buffer (3 column volumes). The residual protein was eluted with the same buffer containing 1 M NaCl. Fractions with high mutanase activity were pooled and concentrated. After the procedure was repeated 12 times, the pooled fractions were concentrated and placed in 10 mM Tris-HCl, pH 8.0. The mutanase was further purified on a HiLoad Q-Sepharose column (50 ml, 2.6 × 10 cm) equilibrated with 10 mM Tris-HCl, pH 8.0, and eluted with a linear gradient from 0 to 50 mM NaCl in 12 column volumes. Fractions with high mutanase activity were pooled and concentrated in an Amicon cell equipped with a 10-kDa cut-off membrane. Finally, the mutanase preparation was dialyzed extensively against 10 mM sodium phosphate, pH 7.0. SDS-PAGE[1] gave one single band at 75 kDa (data not shown).

Carbohydrate composition analysis was performed on lyophilized samples which were hydrolyzed *in vacuo* in sealed glass tubes using 100 μl of 2 M trifluoroacetic acid for 1 h and 4 h at 100 °C. Monosaccharides were separated by high performance anion exchange chromatography using a Dionex Carbopac PA1 column eluted with 16 mM NaOH and detected by pulsed amperometric detection.

The mutanase mass was measured using matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-MS) (VG Analytical). Typically 2 μl of sample were mixed with 2 μl of saturated matrix solution (α-cyano-4-hydroxycinnamic acid in 0.1% trifluoroacetic acid:acetonitrile (70:30)) and 2 μl of the mixture were deposited on the target plate. After evaporation of the solvent, the samples were introduced in the spectrometer. They were desorbed and ionized by 4-ns laser pulses (337 nm) and subjected to an accelerating voltage of 25 kV. Ions were detected by a microchannel plate set at 1850 V.

Generation of a cDNA Probe for the T. harzianum Mutanase Using Reverse Transcriptase PCR—*T. harzianum* was cultivated as described (8). A 2-liter sample was taken after 4 days of growth at 30 °C, and the

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™ */EBI Data Bank with accession number(s) AF214480 (T. harzianum* mutanase gene sequence previously listed as Geneseq™ accession number V12368) and AF214481 (*P. purpurogenum* mutanase gene sequence; previously listed as Geneseq™ accession number V81911).

§ To whom correspondence should be addressed: Novo Nordisk A/S, Novo Allé, Bagsværd DK-2880, Denmark. E-mail: ccf@novo.dk; Tel.: 45-4442-1406; Fax: 45-4442-2202.

|| Present address: Dept. of Microbiology, Dartmouth College, Dartmouth, NH 03755.

** Present address: ProFound-Pharmaceuticals A/S, Roennegade 2, 2100 Copenhagen, Denmark.

[1] The abbreviations used are: PAGE, polyacrylamide gel electrophoresis; MALDI-MS, matrix-assisted laser desorption ionization-mass spectrometry; PCR, polymerase chain reaction; bp, base pair(s); CAPS, 3-(cyclohexylamino)propanesulfonic acid; nt, nucleotide(s); MU, mutanase unit; ORF, open reading frame.

mycelium was collected, frozen in liquid N_2, and stored at −80 °C. First-strand cDNA was synthesized from 5 μg of *T. harzianum* poly(A)$^+$ RNA as described earlier (9). A 387-bp fragment of the *T. harzianum* mutanase cDNA (10) was amplified using two mutanase-specific primers (100 pmol each): forward (5′-ACTAAGCTTTATGTTCAAAATGAGCA-3′) and reverse (5′-ACACTCTAGAACATATGGGTTGAAGTTGT-3′), a DNA thermal cycler (Landgraf, Germany) and 2.5 units of *Taq* polymerase (Perkin-Elmer Cetus). Initially, two cycles of PCR were done using a cycle profile of denaturation at 94 °C for 1 min, annealing at 45 °C for 2 min, and extension at 72 °C for 3 min, then the annealing temperature was increased to 55 °C and 30 additional cycles were performed. The PCR fragment of interest was subcloned into pUC18 vector and sequenced as described previously (9)

Construction and Screening of the T. harzianum cDNA Library—Total RNA was prepared from frozen, powdered mycelium of *T. harzianum* by extraction with guanidinium thiocyanate followed by ultracentrifugation through a 5.7 M CsCl cushion (11). The poly(A)$^+$ RNA was isolated by oligo(dT)-cellulose affinity chromatography (12). Double-stranded cDNA was synthesized from 5 μg of *T. harzianum* poly(A)$^+$ RNA as described earlier (13), except that 25 ng of random hexanucleotide primers (Life Technologies, Inc.) were included in the first strand synthesis. A cDNA library, consisting of 1.5×10^6 independent clones was constructed in the yeast expression vector pYES 2.0 (Invitrogen) as described (13), and screened by colony hybridization (14) using a random-primed (15) ^{32}P-labeled ($>1 \times 10^9$ cpm/μg) mutanase cDNA fragment as a probe. The hybridizations were carried out in 2 × SSC, 5 × Denhardt's solution (14), 0.5% (w/v) SDS, 100 μg/ml denatured salmon sperm DNA for 24 h at 65 °C followed by washes in 2 × SSC (2 × 15 min), 2 × SSC, 0.5% SDS (15 min), 0.2 × SSC, 0.5% SDS (15 min), and finally in 2 × SSC (2 × 15 min) at 65 °C.

Cloning of P. purpurogenum Mutanase Gene—Total cellular DNA was isolated from *P. purpurogenum* cells by a previously described method (16), and used for construction of genomic DNA libraries in the bacteriophage λ-ZipLox cloning system (Life Technologies Inc., Gaithersburg, MD) (17). Approximately 45,000 plaques from the library were screened by plaque hybridization (18) with a radiolabeled *T. harzianum* mutanase probe fragment using moderate stringency conditions (5 × SSPE, 35% formamide (v/v), 0.3% SDS, 200 μg/ml denatured and sheared salmon testes DNA; hybridization temperature 45 °C. Membranes were washed once in 0.2 × SSPE with 0.1% SDS at 45 °C followed by two washes in 0.2 × SSPE (no SDS) at the same temperature.) Plaques which gave hybridization signals were purified twice on *Escherichia coli* Y1090ZL cells, and the mutanase clones were subsequently excised from the λ-ZipLox vector as pZL1-derivatives (19). One such clone, designated pZL-Pp6A, was selected for further study.

DNA Sequence Analysis—DNA sequencing was done with an Applied Biosystems Model 373A Automated DNA Sequencer (Applied Biosystems, Inc., Foster City, CA) using a combination of shotgun DNA sequencing (20) and the primer walking technique with dye-terminator chemistry (21).

Construction of T. harzianum Mutanase Expression Vector—The *T. harzianum* mutanase cDNA fragment was inserted in a two-step cloning procedure into an *A. oryzae* expression vector, pMHan37 (kindly provided by I. G. Clausen, Novo Nordisk A/S), which contains the *A. nidulans amdS* gene as a selectable marker, pUC plasmid sequences for replication *in E. coli*, an α-amylase gene promoter from *A. oryzae*, and the *A. niger* glucoamylase (*glaA*) terminator (22). In the first step, pMHan37 was linearizd with the restriction enzymes *Eco*RI and *Xho*I. This fragment was ligated with the following three segments: 1) a 618-nt fragment of the α-amylase promoter sequence bordered by an *Eco*RI site at the 5′ end and a *Bam*HI site at the 3′ end; 2) linker number 1 listed below which has a *Bam*HI site at the 5′ end and a *Nar*I site at the 3′ end. This linker includes the Met start codon and 12 amino acids of the mutanase signal sequence; and 3) a 68-nt *NarI/XhoI* fragment from the mutanase cDNA clone containing amino acids 12 to 34 of the mutanase gene. The resulting plasmid pJW99 contains the α-amylase promoter immediately upstream from the first 34 amino acids of the mutanase gene, followed by the *A. niger* glucoamylase terminator. To complete the expression vector the mutanase cDNA fragment was cleaved with *Xho*I and *Sph*I giving a 1790-nt fragment encoding amino acids 35–598. This fragment was ligated with pJW99 that had been linearized with *Xho*I plus *Xba*I and linker number 2, yielding the vector pMT1802, which contains the entire mutanase coding region under the transcriptional control of the *A. oryzae* α-amylase promoter and *A. niger* glucoamylase terminator. Plasmid pMT1796 is identical to pMT1802 except that Glu-35 of the mutanase protein has been changed to Lys-35 by replacing the *XhoI/KpnI* fragment of pMT1802 with a PCR-amplified fragment containing this mutation. This PCR fragment was created in a two-step procedure as reported in Ref. 23 using the following primers: Primer 1 (nt 2761, 5′-CAGCGTCCACATCACGAGC, nt 2779) and Primer 2 (nt 3306, 5′-CAAGAAGCACGTTTCTCAGAGACCG, nt 3281); Primer 3 (nt 3281 5′-CGGTCTCTGAGAAACGTGCTTCTTC, nt 3306) and Primer 4 (nt 4276, 5′-GCCACTTCCGTTATTAGCC, nt 4257); nucleotide numbers refer to the pMT1802 plasmid.

```
GATCCTCACA  ATG  TTG  GGC  GTT  GTC  CGC  CGT  CTA  GGC  CTA  GG
    GAGTGT  TAC  AAC  CCG  CAA  CAG  GCT  GCA  GAT  CCG  GAT  CCG  C
            Met  Leu  Gly  Val  Val  Arg  Arg  Leu  Gly  Leu  Gly
```

LINKER 1

```
          C  CAA  TAC  TGT  TAG  T
    GT  ACG  GTT  ATG  ACA  ATC  AGATC
   Ala  Cys  Gln  Tyr  Cys  ***
```

LINKER 2

Expression of Recombinant T. harzianum Mutanase in A. oryzae—The *A. oryzae* host strain JaL125 was transformed using a polyethylene glycol-mediated protocol (24) and a DNA mixture containing 0.5 μg of a plasmid encoding the gene that confers resistance to the herbicide Basta (25) and 8.0 μg of the expression vector pMT1796. Transformants were selected on minimal plates containing 0.5% Basta and 50 mM urea as a nitrogen source. Each transformant was purified twice on selection media and conidia were harvested. Universal containers (20 ml, Nunc, catalog number 364211) containing 10 ml of YPM (2% maltose, 1% bactopeptone bactopeptone, and 0.5% yeast extract) were inoculated with spores from the transformants and incubated 5 days with shaking at 30 °C. Culture supernatants were harvested after 5 days growth and assayed for the recombinant mutanase.

Expression of P. purpurogenum Mutanase in A. oryzae—Two synthetic oligonucleotide primers were designed to amplify the *P. purpurogenum* mutanase gene from plasmid pZL-Pp6A, 5′-cccatttaaatATGAAAGTCTCCAGTGCCTTC and 5′-cccttaattaaTTAGCTCTCTACTTGACAAGC (capital letters correspond to the sequence present in the mutanase coding region). One hundred picomoles of each primer was used in a PCR reaction containing 52 ng of plasmid DNA, 1× *Pwo* polymerase buffer (Roche Molecular Biochemicals, Indianapolis, IN), 1 mM each dATP, dTTP, dGTP, dCTP, and 2.5 units of *Pwo* polymerase (Roche Molecular Biochemicals). The PCR conditions were 95 °C 3 min, 25× (95 °C 1 min, 60 °C 1 min, 72 °C 1.5 min), 72 °C 5 min. The amplified 2.2-kilobase DNA fragment was purified by gel electrophoresis and cut with restriction endonucleases *Swa*I and *Pac*I (using conditions specified by the manufacturers). The fragment was cloned into plasmid pBANe6 (26) that had been previously cut with *Swa*I and *Pac*I and the resultant expression plasmid was named pJeRS35. This vector was introduced into *A. oryzae* host strain JaL142 using a standard protoplast transformation procedure (24) and 40 transformants were selected by their ability to grow on COVE medium using acetamide as sole nitrogen source. The transformants were grown in 20 ml of MY50N media (MY50N in g/liter: Nutriose (Roquette), 62; $MgSO_4 \cdot 7H_2O$, 2.0; KH_2PO_4, 2.0; citric acid, 4.0; yeast extract, 8.0; urea, 2.0; trace metals, 0.5 ml; pH 6.0, and then add $CaCl_2$, 0.1) in shaker flasks for 3 days at 34 °C with agitation. Mutan assay plates were prepared by blending a suspension of 1% (v/w) mutan, 1% agarose in 0.1 M sodium acetate buffer, pH 5.5, for 20 min at 4 °C. The agarose was melted by heating and 150-mm Petri plates were poured. After solidification, small wells (about 40 μl equivalent volume) were punched in the plates. To screen the transformants for ability to secrete mutanase, 35 μl of centrifuged culture broth from each transformant (and one untransformed control) were pipetted into the wells and the plates were incubated at 37 °C. Mutanase activity in the broth samples caused formation of clearing

zones around the wells.

Preparation of Mutan—Mutan was prepared by growing *S. mutans* CBS 350.71 at 37 °C, pH 6.5 (kept constant at a stirring rate of 75 rpm in a medium comprised of the following components: NZ-Case, 6.5 g/liter; yeast extract, 6 g/liter; $(NH_4)_2SO_4$, 20 g/liter; K_3PO_4, 3 g/liter; glucose, 50 g/liter; pluronic PE6100, 0.1%). After 35 h, sucrose was added to a final concentration of 60 g/liter to induce glucosyltransferase. The total fermentation time was 75 h. The supernatant from this fermentation was centrifuged and filtered (sterile). Sucrose was added to the supernatant to a final concentration of 5% (pH was adjusted to pH 7.0 with acetic acid) and the solution was stirred overnight at 37 °C. The solution was filtered and the insoluble mutan harvested on a Propex 23 filter (Scapa Filtration) and washed with deionized water containing 1% sodium benzoate, pH 5 (adjusted with acetic acid). Finally, the insoluble mutan was lyophilized and ground.

Enzyme Assays—The production of soluble reducing sugars released from mutan was employed as a measure of enzyme activity. First, 0.1 ml of 5% mutan in 50 mM sodium acetate (allowed to swell at least for 1 h), pH 5.5, was added to 0.3 ml of enzyme sample (diluted in water) in a round-bottomed Eppendorf vial to ensure sufficient agitation and incubated for 15 min at 40 °C while shaking vigorously. The reaction was terminated by adding 0.1 ml of 0.4 M NaOH and the samples were centrifuged for 5 min at 14,000 × *g* and filtered through 0.45-μm HV-filters (Millipore). To each filtrate (100 μl) in Eppendorf vials 750 μl of ferricyanide reagent (0.4 g/liter $K_3Fe(CN)_6$, 20 g/liter Na_2CO_3) was added and incubated 15 min at 85 °C. After allowing the samples to cool, the decrease in absorbance at 420 nm was measured. A dilution series of glucose was included as a standard. Proper controls (substrate and enzyme blanks) were always included. One mutanase unit (MU) was defined as the amount of enzyme releasing 1 μmol of reducing sugar per minute at pH 5.5 and 40 °C. Temperature profiles were obtained by incubating the assay mixture (50 mM sodium acetate, pH 5.5) at various temperatures. The pH profiles were obtained by suspending the mutan in 50 mM buffer at various pH (glycine-HCl, pH 3–3.5; sodium acetate, pH 4–5.5; and sodium phosphate, pH 6–7.5).

Purification of Recombinant T. harzianum Mutanase—The fermentation broth (700 ml) containing 15.4 MU/ml was filtered using GF/A (Whatmann) and HV 0.45-μm (Millipore) filters and concentrated on a Filtron concentrator equipped with a 10 kDa cut-off membrane. The pH was adjusted to 4.7 (conductivity approximately 300 microsiemens/cm), and the broth was loaded onto an S-Sepharose column (XK 50/22, Amersham Pharmacia Biotech) equilibrated in 10 mM sodium acetate, pH 4.7. The mutanase was eluted in a linear NaCl gradient. Fractions containing mutanase activity were pooled and concentrated on an Amicon Cell (YM10) and loaded onto a HiLoad Q-Sepharose column (Amersham Pharmacia Biotech) equilibrated in 10 mM Tris-HCl, pH 8.0 (approximately 600 μS/cm), in three rounds. The mutanase was eluted in a linear gradient of NaCl. Pooled fractions (according to activity/purity) were concentrated and further purified by gel filtration on a Superdex 75 (16/60) column (Amersham Pharmacia Biotech) in 0.1 M sodium acetate, pH 6.0.

Purification of Recombinant P. purpurogenum Mutanase—The fermentation broth (780 ml) containing 2.2 MU/ml was filtered (0.45 μm; HV Millipore) and mixed with 15.6 g of mutan, washed in 0.1 M sodium acetate, pH 5.5, to provide a 2% solution. The pH was adjusted to 5.5 and the suspension was allowed to stand at 4 °C for 1 h while stirring. The suspension was then filtered on a sintered glass filter funnel and the mutan was washed four times with 0.1 M sodium acetate, pH 5.5 (total volume: 1110 ml), and then six times with Milli Q-filtered deionized water (total volume, 1250 ml); after each washing step the suspension was filtered. The mutanase eluted during the washing with water. These filtrates were pooled, filtered (0.7 μm, Whatman), concentrated on a Filtron concentrator equipped with a 10 kDa cut-off membrane, and further concentrated to 25 ml on an Amicon cell (YM10 membrane).

Preparation of Binding Isotherms—Equilibrium binding was ascertained with 10 mg/ml mutan incubated at 4 °C with 0.5 MU mutanase in 10 mM Britton-Robinson buffer, pH 7. At various time points, samples were taken and filtered (0.45-μm HV, Millipore) prior to measuring the activity. Binding isotherms were obtained by incubating various concentrations of purified mutanase in a 0.2% suspension of mutan in 0.1 M sodium phosphate, pH 7, for 1 h at 4 °C while stirring. The mutan was rinsed in buffer prior to use. Samples were then centrifuged for 10 min at 15,000 × *g* and the amount of enzyme left in the supernatant determined by fluorescence spectrometry (Perkin-Elmer LS50) with excitation at 280 nm and emission at 345 nm. A fluorescence standard curve of the enzyme diluted in buffer was always included. Alternatively, the activity was measured in the supernatant and compared with the control. The data was fitted using the simple Langmuir theory for adsorption to a surface: $A = (A_{max} \times E_{free})/(K_D + E_{free})$, where A is the adsorbed protein, A_{max} is the maximum amount of protein which can be adsorbed to the surface, E_{free} is the free protein and K_D the equilibrium constant for $ES \leftrightarrow E + S$ (27).

SDS-PAGE—SDS-PAGE was done with 4–20 or 8–16% gradient gels (Novex) according to the manufacturer's instructions.

Differential Scanning Calorimetry—Samples for DSC were desalted into the appropriate buffer using NAP-5 columns from Amersham Pharmacia Biotech. Final enzyme concentrations were in the range from 2 to 3 mg/ml. Samples were scanned from 20 to 90 °C using a scan rate of 90°/h at the MC-2 (MicroCal).

Protein Sequencing—NH_2-terminal amino acid sequencing was done using an Applied Biosystems 473A protein sequencer according to the manufacturer's instructions.

Isolation and Mutan Binding Activity of the COOH-terminal Domain from the T. harzianum Mutanase—Mutanase was incubated for 2.5 h at 30 °C with chymotrypsin (Roche Molecular Biochemicals) in a ratio of 100:1 (mutanase:chymotrypsin, w/w) in 50 mM NH_4HCO_3. The digest was investigated on SDS-PAGE and the 41-kDa band observed was electroblotted from SDS-PAGE onto a Millipore Immobilin P^{SQ} polyvinylidene difluoride membrane in 10 mM CAPS, 6% methanol at 175 mA for 3 h and subjected to NH_2-terminal amino acid sequencing, revealing a sequence of SLTIGL- corresponding to proteolytic cleavage after amino acid residue Phe-473. A 50-μl sample of the digest was incubated for 30 min at room temperature with 50 μl of 2.5% mutan suspension. The sample was centrifuged for 2 min at 15,000 × *g*. A 30-μl volume of supernatant was then analyzed by SDS-PAGE (Novex 4–20%). Controls without mutan were included.

RESULTS

Wild-type T. harzianum Mutanase—Purified wild-type *T. harzianum* mutanase displayed a molecular mass of 75 kDa both in SDS-PAGE and MALDI-MS. Carbohydrate composition analysis revealed only glucose and mannose but no *N*-acetylglucosamine, indicating *O*-glycosylation. The amount of glucose and mannose (18 and 32 mol/mol enzyme, respectively) accounts for over 8 kDa which, added to the theoretical mass (63.8 kDa), gives a molecular mass of about 72 kDa in close agreement with the 75 kDa measured by MALDI-MS and SDS-PAGE.

Isolation and Characterization of cDNA Clones Encoding the Mutanase from T. harzianum—To obtain a cDNA probe for the *T. harzianum* mutanase, two oligonucleotides based on a genomic mutanase clone from *T. harzianum* (10) were designed. These primers were used to amplify a mutanase cDNA fragment from *T. harzianum* first-strand cDNA employing the PCR technique (28). Sequencing of the subcloned PCR fragment revealed a 387-bp cDNA with an open reading frame of 129 amino acids. In addition to the primer-encoded residues, the ORF was identical to the corresponding region in the *T. harzianum* mutanase amino acid sequence (10), confirming that the PCR had specifically amplified the desired cDNA species. Approximately 10,000 colonies from a *T. harzianum* cDNA library in *E. coli* were screened using the mutanase-specific PCR product as a probe. This yielded 12 positive clones with inserts ranging from 0.8 to 2.0 kilobase. These were further analyzed by sequencing the ends of the cDNAs with forward and reverse pYES polylinker primers, and determining the nucleotide sequence of the longest cDNA from both strands with synthetic oligonucleotide primers. The nucleotide sequence and the deduced amino acid sequence of the mutanase cDNA from *T. harzianum* are presented in Fig. 1. The 2062-bp cDNA clone contains a 1905-bp open reading frame initiating with an ATG codon at nucleotide position 29 and terminating with a TAG stop codon at nucleotide position 1931, thus predicting a 634-residue polypeptide. The open reading frame is preceded by a 28-bp 5′-noncoding region and followed by a 119-bp 3′-noncoding region and a poly(A) tail.

Cloning of P. purpurogenum Mutanase—Southern blotting experiments indicated that the *T. harzianum* mutanase cDNA could be used as a probe to identify mutanase gene-specific

A

caagcagaatccatctaaaacaccctcaATGTTGGGCGTTGTCCGCCGCCTCGGGCTCGGCGCCCTTGCTGCCGCAGCTCTGTCTTCTCT
90
M L G V V R R L G L G A L A A A A L S S L

CGGCAGTGCCGCTCCCGCCAATGTTGCTATTCGGTCTCTCGAGGAACGTGCTTCTTCTGCTGACCGTCTCGTATTCTGTCACTTCATGAT
180
G S A A P A N V A I R S L E E R A S S A D R L V F C H F M I

TGGTATTGTTGGTGACCGTGGCAGCTCAGCAGACTATGATGATGACATGCAACGTGCCAAAGCCGCTGGCATTGACGCATTCGCTCTGAA
270
G I V G D R G S S A D Y D D D M Q R A K A A G I D A F A L N

CATCGGCGTTGACGGCTATACCGACCAGCAACTCGGGTATGCCTATGACTCTGCCGACCGTAATGGCATGAAAGTCTTCATTTCATTCGA
360
I G V D G Y T D Q Q L G Y A Y D S A D R N G M K V F I S F D

TTTCAACTGGTGGAGCCCCGGTAATGCAGTTGGTGTTGGCCAGAAGATTGCGCAGTATGCCAGCCGTCCCGCCCAGCTGTATGTTGACAA
450
F N W W S P G N A V G V G Q K I A Q Y A S R P A Q L Y V D N

CCGGCCATTCGCCTCTTCCTTCGCTGGTGACGGTTTGGATGTAAATGCGTTGCGCTCTGCTGCAGGCTCCAACGTTTACTTTGTGCCCAA
540
R P F A S S F A G D G L D V N A L R S A A G S N V Y F V P N

CTTCCACCCTGGTCAATCTTCCCCCTCCAACATTGATGGCGCCCTCAACTGGATGGCCTGGGATAATGATGGAAACAACAAGGCACCCAA
630
F H P G Q S S P S N I D G A L N W M A W D N D G N N K A P K

GCCGGGCCAGACTGTCACGGTGGCAGACGGTGACAACGCTTACAAGAATTGGTTGGGTGGCAAGCCTTACCTAGCGCCTGTCTCCCCTTG
720
P G Q T V T V A D G D N A Y K N W L G G K P Y L A P V S P W

GTTTTTCACCCATTTTGGCCCTGAAGTTTCATATTCCAAGAACTGGGTCTTCCCAGGTGGTCCTCTGATCTATAACCGGTGGCAACAGGT
810
F F T H F G P E V S Y S K N W V F P G G P L I Y N R W Q Q V

CTTGCAGCAGGGCTTCCCCATGGTTGAGATTGTTACCTGGAATGACTACGGCGAGTCTCACTACGTCGGTCCTCTGAAGTCTAAGCATTT
900
L Q Q G F P M V E I V T W N D Y G E S H Y V G P L K S K H F

CGATGATGGCAACTCCAAATGGGTCAATGATATGCCCCATGATGGATTCTTGGATCTTTCAAAGCCGTTTATTGCTGCATATAAGAACAG
990
D D G N S K W V N D M P H D G F L D L S K P F I A A Y K N R

GGATACTGATATATCTAAGTATGTTCAAAATGAGCAGCTTGTTTACTGGTACCGCCGCAACTTGAAGGCATTGGACTGCGACGCCACCGA
1080
D T D I S K Y V Q N E Q L V Y W Y R R N L K A L D C D A T D

CACCACCTCTAACCGCCCGGCTAATAACGGAAGTGGCAATTACTTTATGGGACGCCCTGATGGTTGGCAAACTATGGATGATACCGTTTA
1170
T T S N R P A N N G S G N Y F M G R P D G W Q T M D D T V Y

TGTTGCCGCACTTCTCAAGACCGCCGGTAGCGTCACGGTCACGTCTGGCGGCACCACTCAAACGTTCCAGGCCAACGCCGGAGCCAACCT
1260
V A A L L K T A G S V T V T S G G T T Q T F Q A N A G A N L

CTTCCAAATCCCTGCCAGCATCGGCCAGCAAAAGTTTGCTCTAACTCGCAACGGTCAGACCGTCTTTAGCGGAACCTCATTGATGGATAT
1350
F Q I P A S I G Q Q K F A L T R N G Q T V F S G T S L M D I

CACCAACGTTTGCTCTTGCGGTATCTACAATTTCAACCCATATGTTGGCACCATTCCTGCCGGCTTTGACGACCCTCTTCAGGCTGACGG
1440
T N V C S C G I Y N F N P Y V G T I P A G F D D P L Q A D G

TCTTTTCTCTTTGACCATCGGATTGCATGTCACGACTTGTCAGGCCAAGCCATCTCTTGGAACCAACCCTCCTGTCACTTCTGGCCCTGT
1530
L F S L T I G L H V T T C Q A K P S L G T N P P V T S G P V

GTCCTCGCTGCCAGCTTCCTCCACCACCCGCGCATCCTCGCCTCCTGTTTCTTCAACTCGTGTCTCTTCTCCCCCTGTCTCTTCCCCTCC
1620
S S L P A G S T T E A S S P P V S S T R V S S P P V S S P P

B

AGTTTCTCGCACCTCTTCTCCCCCTCCCCCTCCGGCCAGCAGCACGCCGCCATCGGGTCAGGTTTGCGTTGCCGGCACCGTTGCTGACGG
1710
V S E T S S P P P P P A S S T P P S G Q V G V A G T V A D G

CGAGTCCGGCAACTACATCGGCCTGTGCCAATTCAGCTGCAACTACGGTTACTGTCCACCGGGACCGTGTAAGTGCACCGCCTTTGGTGC
1800
E S G N Y I G L C Q F S C N Y G Y C P P G P C K C T A F G A

TCCCATCTCGCCACCGGCAAGCAATGGGCGCAACGGCTGCCCTCTACCGGGAGAAGGCGATGGTTATCTGGGCCTGTGCAGTTTCAGTTG
1890
P I S P P A S N G R N G C P L P G E G D G Y L G L C S F S C

TAACCATAATTACTGCCCGCCAACGGCATGCCAATACTGTTAGgagagagatcaatctcagtatgagtatatggaggctgccgaaggacc
1980
N H N Y C P P T A C Q Y C *

agttagctgttcttatcggcagacgaaacccatagagtaagaagttaaataaaatgcaattaatgtgttgtcaaaaaaaaaa

2062

FIG. 1. **The nucleotide sequence and the deduced amino acid sequence of the α1,3-glucanase (mutanase) cDNA from *T. harzianum*.** The signal peptide and propeptide region are *underlined*, and the NH_2-terminal residues determined from the purified, recombinant *T. harzianum* mutanase are indicated by *double underlines*. The putative linker region (rich in Ser, Pro, and Thr) flanked by Cys residues at positions 484 and 553 is highlighted in *gray*. Noncoding sequences are in *lowercase letters*. This sequence has been deposited in the Geneseq™ data base with the accession number V12368.

fragments in *P. purpurogenum* genomic DNA (data not shown). Consequently, a genomic library was constructed from *P. purpurogenum* cellular DNA using the bacteriophage vector λ-ZipLox. Approximately 45,000 plaques from this library were screened by hybridization using a segment of the *T. harzianum* mutanase cDNA as the probe. Eighteen positive clones which hybridized strongly to the probe were picked and 10 were plaque-purified (18) and excised from the λ cloning vector using the *in vivo* excision protocol (19). Preliminary restriction mapping on one of the pZL1-mutanase clones (designated pZL-

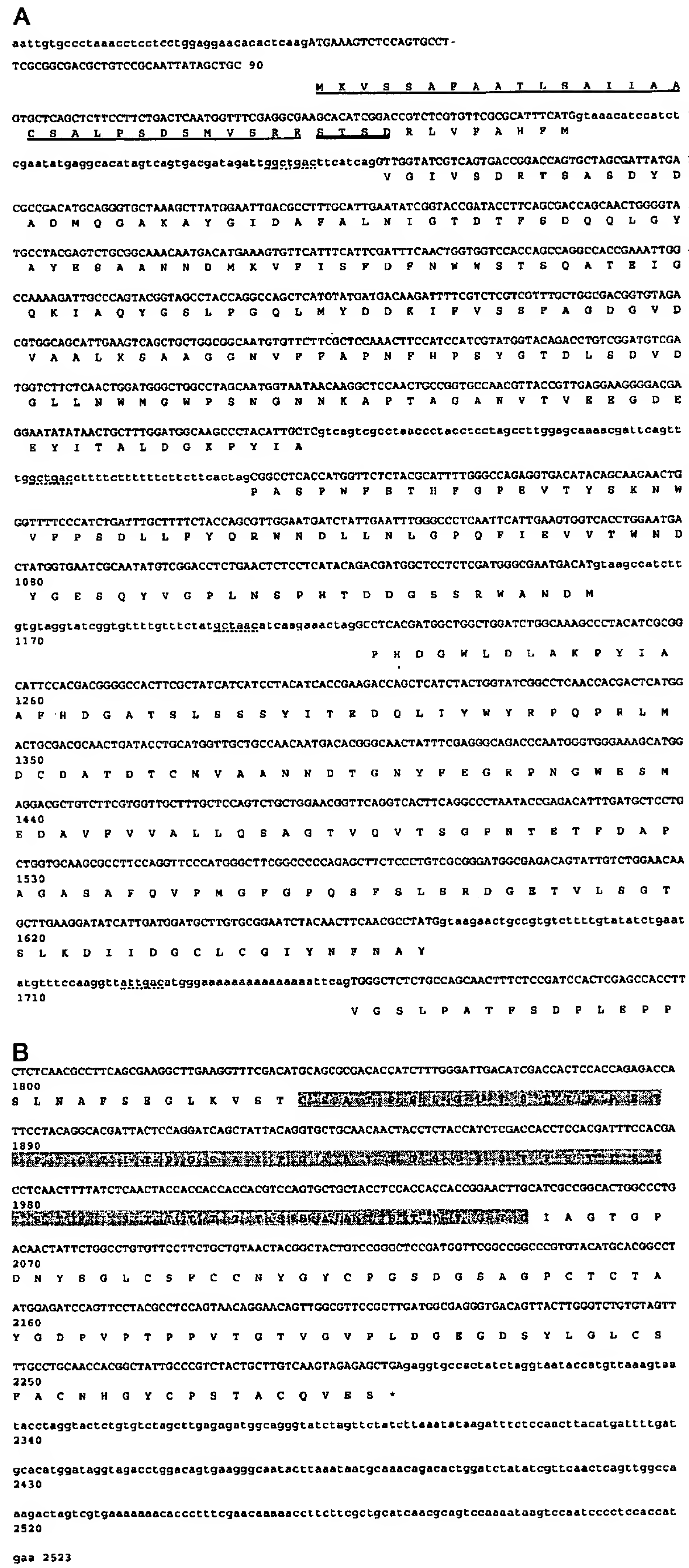


FIG. 2. **DNA sequence and deduced amino acid sequence of *P. purpurogenum* mutanase gene.** The signal peptide and propeptide region are *underlined*, and the NH_2-terminal residues determined from the purified, recombinant *P. purpurogenum* mutanase are indicated by *double underlines*. The putative linker region (rich in Ser, Pro, and Thr residues) flanked by Cys residues at positions 477 and 547 is highlighted in *gray*. Introns and noncoding regions are indicated by *lowercase letters*. Consensus lariat sequences (PuCTPuAC) with each intron are denoted by a *dashed underline*. This sequence has been deposited in the Geneseq™ data base with the accession number T89024.

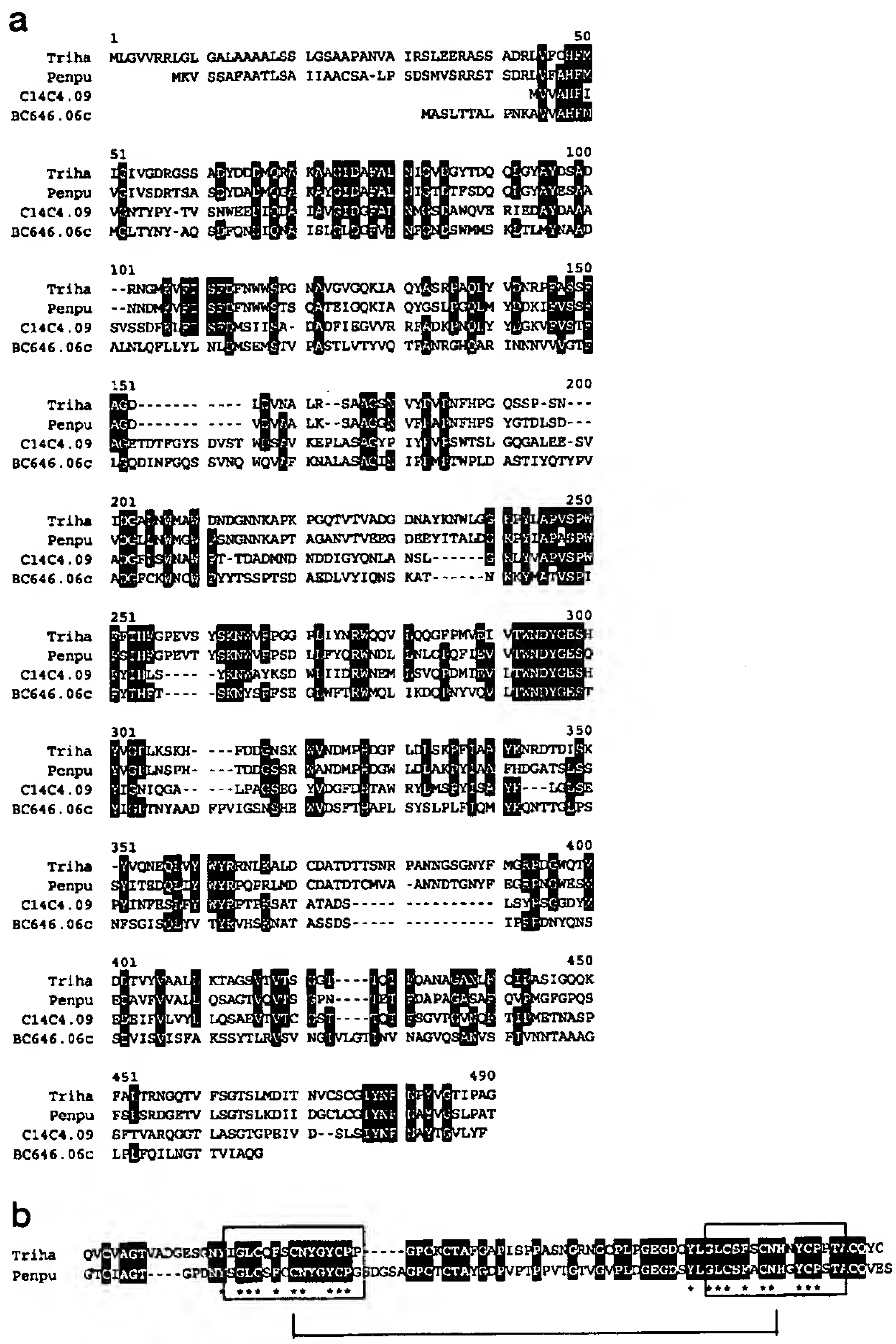


FIG. 3. *a*, sequence alignment of the catalytic domains of *T. harzianum* (Triha) and *P. purpurogenum* (Penpu) mutanases (Geneseq™ protein data base accession numbers W44193 and W32213, respectively) with the two homologous *S. pombe* ORFs of unknown function (C[14]C4.09 and BC646.06c, GenBank Z98596 and AL035216, respectively). Residues identical in 3 of the 4 sequences are printed in *white* on *black background*. *b*, alignment of the mutan-binding domains of *T. harzianum* (Triha) and *P. purpurogenum* (Penpu) mutanases. Identical residues are printed in *white* on *black background*. The two regions of internal similarity are *boxed*.

Pp6A) revealed that the region which hybridized to the *T. harzianum* mutanase cDNA was localized near one end of a 3.6-kilobase genomic DNA insert (not shown). DNA sequencing of a portion of this segment showed an open reading frame with clear homology to the *T. harzianum* mutanase cDNA and its deduced amino acid sequence (Fig. 2). The positions of introns and exons within the *P. purpurogenum* mutanase gene were assigned based on alignments of the deduced amino acid sequences to the corresponding *T. harzianum* mutanase gene product. On the basis of this comparison, the *P. purpurogenum* mutanase gene is composed of five exons (126, 532, 226, 461, and 548 bp) which are punctuated by four small introns (63, 81, 58, and 78 bp). These appear to be typical fungal introns with respect to size and composition in that all contain consensus splice donor and acceptor sequences as well as the consensus lariat sequence (PuCTPuAC) near the 3′ end of each intervening sequence (29).

Comparison of Trichoderma and Penicillium Mutanase Primary Structures—The signal peptide and propeptide portions of *P. purpurogenum* mutanase and *T. harzianum* mutanase share little amino acid sequence similarity; however, the mature polypeptides (after removal of signal and propeptides) are approximately 53% identical. The regions of greatest identity are located in the NH_2-terminal portion (residues 42 through 491; *T. harzianum* numbering) and over approximately the last 70 residues of these two proteins where 60 and 63% identity is observed, respectively. In both mutanases the NH_2-terminal and COOH-terminal domains are separated by a Pro-Ser-Thr-rich linker region. Remarkably, the Pro-Ser-Thr-rich region of *P. purpurogenum* mutanase (residues 475 through 547) is composed of 69% Pro, Ser, and Thr, and is bordered roughly by Cys residues at positions 477 and 547. As the two mutanases appeared to have a modular structure, sequence comparisons using the BLAST algorithm to search the non-redundant GenBank CDS translations on the NCBI server (30) were therefore conducted on each domain separately. BLAST searches using the NH_2-terminal domains did not produce any hits with known glycosidases, however, two ORFs of unknown function in *Schizosaccharomyces pombe* (C[14]C4.09 and BC646.06c, GenBank Z98596 and AL035216, respectively) were picked with highly significant scores (*E* values ranging from 6×10^{-50} to 10^{-26}) suggesting that these ORFs encode similar glycosidases.

FIG. 4. **Schematic maps of the vectors pMT1796 and pJeRS35 for the heterologous expression of *T. harzianum* and *P. purpurogenum* mutanases in *A. oryzae* (not to scale).**

An alignment of the sequences of the catalytic domains of the two mutanases with the two ORFs of *S. pombe* is shown in Fig. 3a. BLAST searches conducted with the COOH-terminal domains of the mutanases also failed to produce any significant hit in GenBank. Within the COOH-terminal domains of the two mutanases, two short regions display intriguing similarity (10 residues conserved out of 15) suggesting the existence of an internal duplication (Fig. 3*b*).

Heterologous Expression of T. harzianum Mutanase in A. oryzae—The *T. harzianum* mutanase coding region was amplified by PCR, and the amplicon was inserted into an *Aspergillus* expression vector so that the gene was under the control of an *A. orzyae* α-amylase gene promoter and an *A. niger glaA* terminator. The resulting expression construct, pMTH1802, was further modified by changing aa 35 from Glu to Lys resulting in the presence of a dibasic (KEX2-type) processing site at the amino terminus of the mature mutanase protein. This new expression vector, pMT1796 (Fig. 4*a*), was used to transform an *A. oryzae* strain, and 25 independent transformants were isolated. Mycelia from each transformant was used to inoculate 20-ml culture tubes containing 10 ml of YPM media and cultures were grown with shaking for 5 days at 30 °C. SDS-PAGE analysis revealed a dominant 85–90-kDa band indicating that these transformants were indeed expressing the recombinant mutanase gene.

Heterologous Expression of P. purpurogenum Mutanase—The *P. purpurogenum* mutanase coding region was amplified by PCR using primers that created 5′- and 3′-terminal restriction sites compatible with an *Aspergillus* expression vector. The amplified DNA segment was subsequently inserted into the vector which employed a strong *A. oryzae* α-amylase gene promoter. The resulting plasmid, designated pJeRS35 (Fig. 4*b*), was used to transform an *A. oryzae* recipient strain, and 40 transformants were isolated. Mycelia from each of the transformants were used to inoculate shaker flask cultures that were incubated for 3 days. Using a mutan agar plate assay, 14 of the transformants showed extracellular mutanase activity as indicated by opaque clearing zones (the control showed no clearing zone). Broth samples that were positive in the plate assay were subsequently analyzed by SDS-PAGE. These transformants showed a prominent band at approximately 90 kDa.

Purification of and Molecular Properties of Recombinant Mutanases—Recombinant *T. harzianum* mutanase was purified in a three-step procedure using cation-exchange chromatography, anion-exchange chromatography followed by size exclusion chromatography resulting in a yield of around 24%. The essentially pure mutanase exhibited a molecular mass around 86 kDa (Fig. 5*a*). A rather broad band was observed indicating some heterogeneity and/or heavy glycosylation. The NH_2-terminal amino acid sequence was determined by protein sequencing to be Ala-Ser-Ser- thus predicting a calculated molecular mass of 63.8 kDa for the mature enzyme (Table I). This obser-

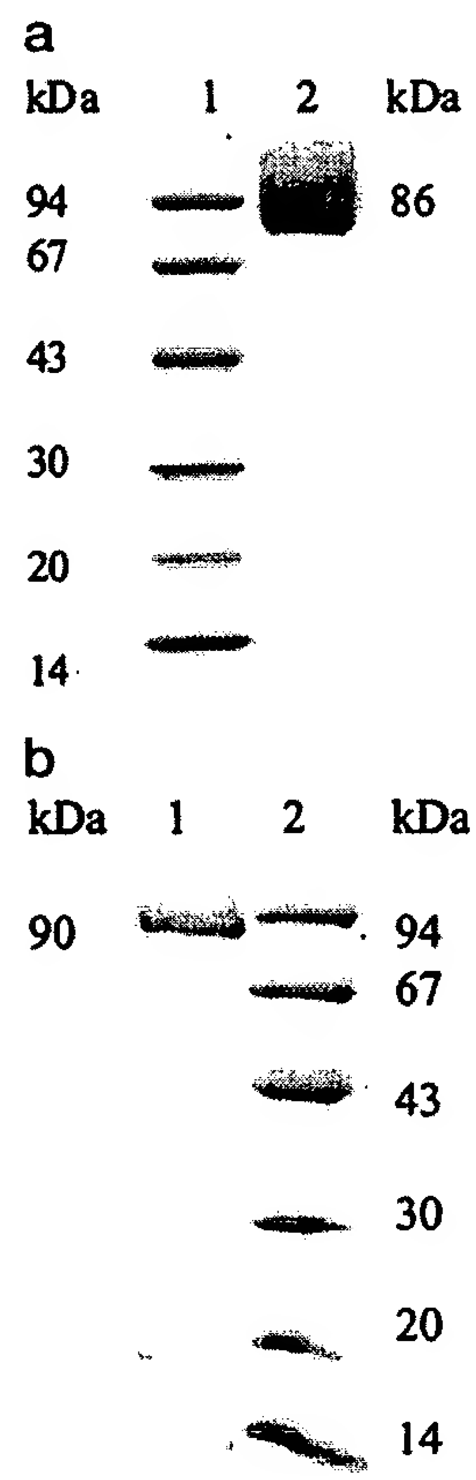


FIG. 5 **SDS-PAGE 4–20% (Novex).** *a*, low molecular mass standard (*lane 1*); purified rec. *T. harzianum* mutanase (*lane 2*). *b*, purified recombinant *P. purpurogenum* mutanase (*lane 1*), low molecular mass standard (*lane 2*).

TABLE I
Molecular properties of purified recombinant mutanases

Enzyme	NH_2-terminal sequence	Starting at residue no.	*m* (calculated)	*m* (SDS)	pI (calculated)
T. harzianum	Ala-Ser-Ser-	38	63,785 Da	86 kDa	5.24
P. purpurogenum	Ser-Thr-Ser-	31	63,557 Da	90 kDa	3.88

vation suggests that the first 37 amino acid residues deduced from the gene sequence function as a secretory signal peptide and propeptide. This is supported by the fact that the NH_2 terminus of the mature mutanase is not preceded by a typical signal peptidase cleavage site (31, Fig. 1) but rather by a cleavage site for a monobasic processing enzyme. Furthermore, the mutanase cDNA encodes an apparent signal sequence of 24 amino acids, with a predicted signal peptidase cleavage site between Ala-24 and Ala-25 in the mutanase precursor (31). A simpler procedure for purification of the recombinant *P. purpurogenum* mutanase was established using the information that a putative COOH-terminal-binding domain is present in the enzyme. The enzyme was adsorbed to insoluble mutan and subsequently eluted in water. This procedure resulted in a 129-fold purification and a yield around 20%. The essentially pure mutanase had a molecular mass of about 90 kDa (Fig. 5*b*). NH_2-terminal amino acid sequencing revealed the following sequence: Ser-Thr-Ser-Asp-Arg-. Thus, the deduced amino acid sequence of the mutanase gene product (Fig. 2) predicts an amino-terminal extension of 30 amino acids which are not present in the mature enzyme and a molecular mass for the mature enzyme of 63.6 kDa (Table I). Based on the rules of von Heijne (31), the first 20 amino acids likely comprise a secretory signal peptide, and the next 10 residues probably represent a propeptide segment which is removed by a subsequent proteolytic cleavage following the dibasic Arg-Arg sequence.

Characterization of the Purified Recombinant Mutanases—The two mutanases showed similar catalytic properties. They both exhibit slightly acidic pH optima in the range from pH 3.5 to 5.0 and pH 3.0 to 4.5 for the *T. harzianum and P. purpurogenum* mutanases, respectively. At pH 5.5 the two enzymes have specific activities of 16 and 12 MU/mg, respectively, on insoluble mutan at 40 °C. Also, the two mutanases have virtually identical temperature optima around 50–55 °C at pH 5.5. This correlates with DSC analysis of the thermal stability of the recombinant *T. harzianum* mutanase which shows a midpoint denaturation temperature (T_M) around 56 °C at pH 5.5 identical to that of the wild-type enzyme.

The binding properties of the two mutanases toward insoluble mutan were investigated at steady state conditions at pH 7 and 4 °C in order to limit hydrolysis. The kinetics of adsorption was followed by taking samples from the supernatant of mutanase incubated with mutan. The equilibrium was reached within 5 min, and then no further net adsorption was observed (data not shown). Varying concentrations of mutanase were incubated for 1 h under the above conditions with mutan and the amount of free mutanase was determined by fluorescence spectroscopy (concentrations verified by activity analysis) and the amount of bound enzyme was calculated. Thus, binding isotherms were generated, and the data fitted using the simple Langmuir model for adsorption to a surface (Fig. 6). Rather strong binding was observed with desorption constants of 0.13 and 0.11 μM for the *T. harzianum* and *P. purpurogenum* mutanases, respectively (Table II). A significant difference is observed in the maximum level of enzyme which can be adsorbed to the insoluble mutan 0.549 *versus* 0.244 μmol of enzyme/g of mutan for the *T. harzianum* and *P. purpurogenum* mutanases, respectively (Table II).

In order to probe the hypothesis that the homologous COOH-

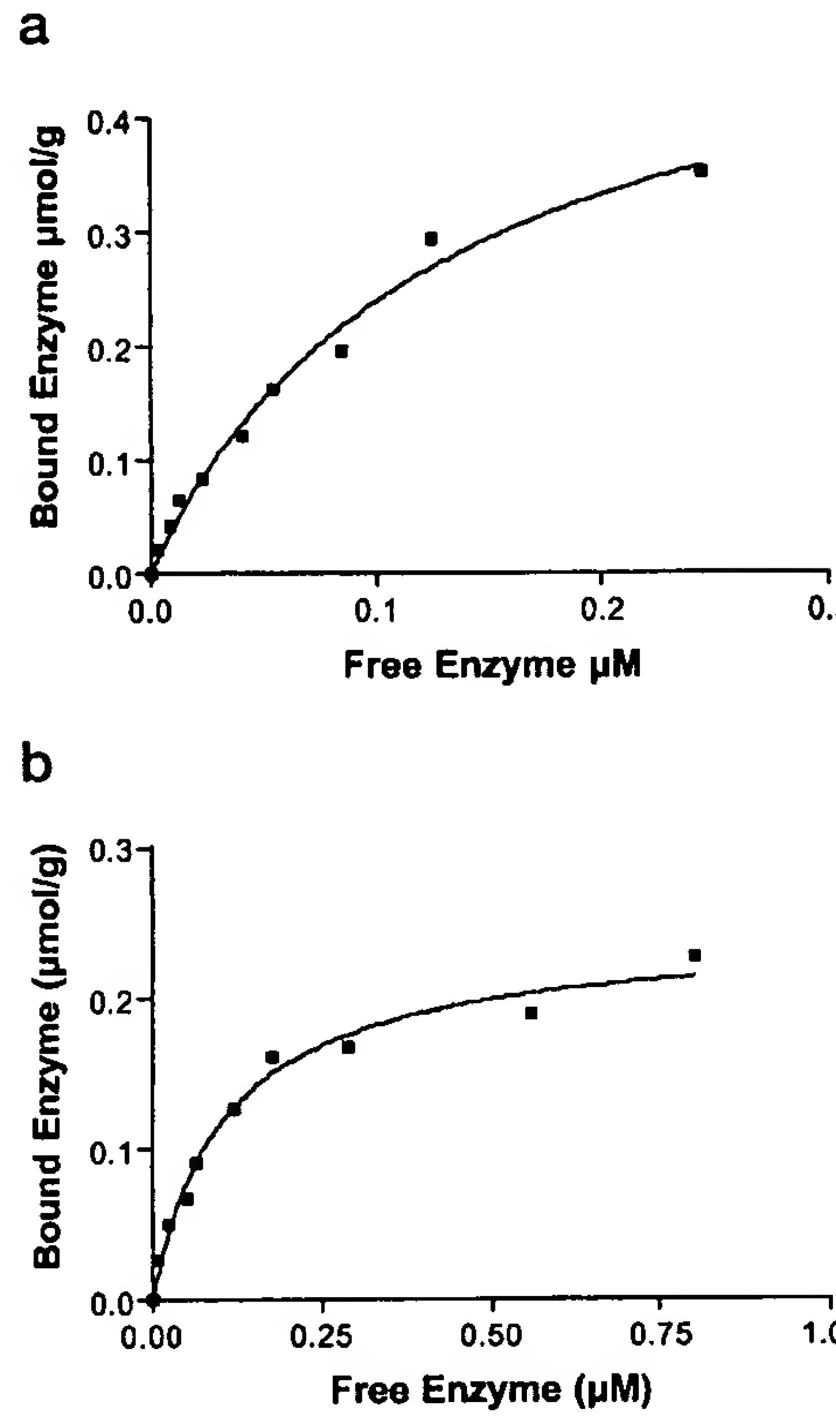


FIG. 6. **Substrate binding isotherms of purified recombinant mutanases; 0.2% mutan in 0.1 M sodium phosphate, pH 7, 4 °C.** *a*, recombinant *T. harzianum* mutanase; *b*, rec. *P. purpurogenum* mutanase.

TABLE II
Substrate binding properties of purified recombinant mutanases

Enzyme	K_o	A_{max}
	μM	*μmol/g mutant*
T. harzianum	0.129 ± 0.021	0.549 ± 0.047
P. purpurogenum	0.111 ± 0.016	0.244 ± 0.012

terminal domain of the two fungal mutanases constitutes a mutant-binding domain, the *T. harzianum* mutanase was subjected to limited proteolysis using chymotrypsin. The protease treatment resulted in a 41-kDa band on SDS-PAGE (Fig. 7), which was NH_2 terminally sequenced after being electroblotted onto a polyvinylidene difluoride membrane revealing the sequence Ser-Leu-Thr-Ile-Gly-Leu- corresponding to proteolytic cleavage between Phe-473 and Ser-474. This strongly suggests that the 41-kDa band corresponds to the linker and the COOH-terminal domain. The chymotrypsin digest of the mutanase was then incubated with 2.5% mutan before centrifuging the sample and loading the supernatant onto SDS-PAGE. From the SDS-PAGE analysis (Fig. 7) it is apparent that the 41-kDa band has been adsorbed to the insoluble mutan since it is no longer present in the supernatant.

DISCUSSION

Nucleic acid sequences encoding extracellular mutanases from the filamentous fungi *T. harzianum* and *P. purpurogenum* were cloned and successfully expressed in *A. oryzae*. The primary translation products of these two DNA sequences appear to be preproenzymes, having both NH_2-terminal signal peptides and propeptides that are removed post-translationally. The two mutanases show deduced amino acid sequence identities of 53% overall. Further analyses of the protein sequences

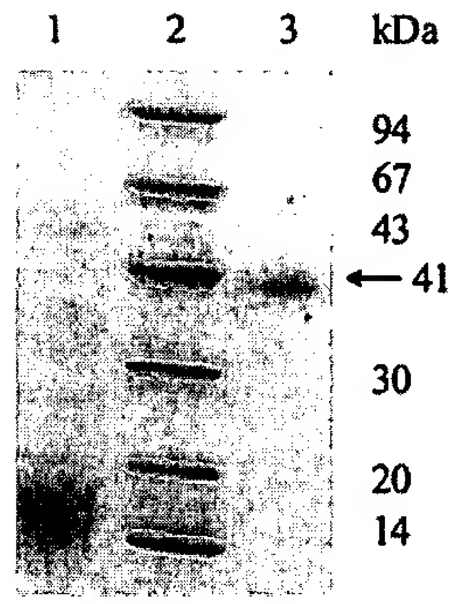


FIG. 7. **SDS-PAGE 4–20% (Novex); 2.5% mutan + chymotrypsin digest of *T. harzianum* mutanase (*lane 1*), low molecular mass standard (*lane 2*), chymotrypsin digest of *T. harzianum* mutanase (*lane 3*).**

reveal stronger similarity between the NH_2-terminal and COOH-terminal parts of the mature enzymes, separated by a less homologous Pro, Ser, and Thr-rich region. Consequently, like many cellulases, glucoamylases, and chitinases, the mature mutanases appear to be made of two distinct domains: a NH_2-terminal catalytic domain, and a putative COOH-terminal polysaccharide-binding domain separated by a Pro-Ser-Thr-rich linker peptide. MALDI-MS and carbohydrate analysis of the wild-type enzyme from *T. harzianum* suggest that the linker region is *O*-glycosylated in a manner similar to the Ser-Thr-rich linker region of *A. niger* glucoamylase (32). The glycosylation is even more pronounced in the recombinant enzymes which display molecular masses of 86 and 90 kDa for the *T. harzianum* and *P. purpurogenum* mutanases, respectively. Fungal polysaccharidases harboring a Pro-Ser-Thr-rich linker separating the catalytic from the carbohydrate-binding domain have long been known to undergo hyperglycosylation upon expression in yeast and other heterologous fungal systems (33).

Experiments showing that the chymotrypsin produced fragment of the *T. harzianum* mutanase adsorbs to mutan gave indirect indication that the COOH-terminal domain of the two fungal mutanases is responsible for binding to insoluble mutan. As a first step in an effort to further characterize the COOH-terminal mutan-binding domain of *T. harzianum* mutanase, two expression plasmids were constructed harboring (i) an internal deletion of the coding region encompassing residues 32–542 (*i.e.* coding for the isolated COOH-terminal binding domain without any linker) and (ii) the NH_2-terminal catalytic domain only. The transformants were tested by immunodiffusion using antibodies raised against the whole mutanase. Whereas the isolated catalytic domain was unaffected by preincubation with mutan, the first transformant became negative upon preincubation with mutan (data to be described elsewhere). The inability of the catalytic domain to bind to mutan was verified by activity anaylsis showing that no activity could be removed from the supernatant by preincubation with insoluble mutan.

Glycoside hydrolases and transglycosidases have been classified in a number of distinct families based on amino acid sequence similarities (34–36). BLAST searches (30) conducted with the NH_2-terminal catalytic domains of the two mutanases described here failed to display any similarity with known glycosidases from previously defined families. Families of glycoside hydrolases being defined with at least two related sequences (34), the two mutanases therefore allow the definition of a new family (designated family 71). Although, sequence similarities between the mutanase catalytic domains and the described ORFs of *S. pombe* are such that it is predictable that all these proteins adopt a similar fold and operate via the same catalytic mechanism using a similar catalytic machinery (37–39), the precise substrate specificity of the *S. pombe* proteins cannot be reliably ascertained as it has been shown that sequence-based families of glycosidases contain enzymes with sometimes widely different substrate specificity (34). Finally, it is worth mentioning that, unlike the two mutanases, none of the two *S. pombe* ORFs carries a COOH-terminal extension suggesting that the encoded proteins are made of a single domain.

Cellulases, xylanases, chitinases, and starch-degrading enzymes have long been recognized to have a modular structure with a catalytic domain carrying one or several ancillary modules whose function is often binding to insoluble polysaccharides (40). The best described of these ancillary modules are probably the cellulose-binding domains which have been classified in several distinct families based on sequence similarities (41, 42). The lack of sequence similarity of the two COOH-terminal domains of the mutanases with any known carbohydrate-binding domains together with their insoluble mutan binding activity allows the definition of a new family of carbohydrate-binding modules.

The pH optimum observed for the two mutanases is not exactly in agreement with the earlier reported pH optimum around pH 6.0 for the *T. harzianum* mutanase (7) but comparable to the pH optimum obtained for the *Trichoderma viride* (43) and the *Penicillium funiculosum* mutanases (3). For comparison, the bacterial mutanases from *Bacillus circulans* (44) and *Streptomyces chartreusis* (45) have slightly higher pH optima than the two fungal mutanases but similar temperature optima. Although, the pH in the oral cavity is around pH 6–7, the slightly acidic pH profile of the two fungal mutanases may be of importance in the application for plaque removal as low pH values have been observed locally in the plaque (46).

The substrate binding constants observed for the two mutanases to insoluble mutan are, although slightly higher, in the range of reported binding constants for cellulase adsorption to insoluble cellulose (27). The difference in the maximum binding capacity observed for the two mutanases may be explained by differences in the batches of mutan used for the experiment as these have been found to vary somewhat in quality/purity. Alternatively, a possible explanation would be that the *T. harzianum* mutanase is capable of dispersing the insoluble mutan (in analogy to cellulose-binding domains and cellulose) to a larger extend than the *P. purpurogenum* mutanase and thus revealing a larger surface area onto which the enzyme can adsorb. The strong adsorption of the fungal mutanases may be beneficial for their application in dentrifice as the enzymes are expected to bind to dental plaque and thus be retained in the oral cavity where it is supposed to exhibit its action in removing the dental plaque.

Acknowledgments—We thank Elizabeth Golightly and Lissi Willum Nielsen for DNA sequencing; Heidi Heinsøe and Maria Juul Holm for skillful technical assistance; Inger Christina Aalvik and Inge Høegh for assistance in vector construction; Birthe Ravn for cultivation of recombinant *A. oryzae* strains; Kim Brown for amino acid sequencing; Jan Lehmbeck for *A. oryzae* host strains and Beth Nelson for the expression vector pBANe6.

REFERENCES

1. Hamada, S., and Slade, H. D. (1980) *Microbiol. Rev.* **44,** 331–384
2. Ginsburg, I. (1982) in *Microbiology* (Braude, A. I., ed) p. 292, W. B. Saunders Co., London
3. Guggenheim, B. (1970) *Helv. Odont. Acta* **14,** 89–108
4. Tsuchiya, R., Fuglsang, C. C., Johansen, C., and Aaslyng, D. (1998) *J. Dent. Res.* **77,** 2713
5. Guggenheim, B., Regolati, B., Schmid, R., and Mühlemann, H. R. (1980) *Caries Res.* **14,** 128–135
6. Kelstrup, J., Holm-Pedersen, P., and Poulsen S. (1978) *Scand. J. Dent. Res.* **86,** 93–102
7. Guggenheim, B., and Haller, R. (1972) *J. Dent. Res.* **51,** 394–402
8. Draborg, H., Kauppinen, S., Dalbøge, H., and Christgau, S. (1995) *Biochem.*

Mol. Biol. Int. **36,** 781–791

9. Siggaard-Andersen, M., Kauppinen, S., and von Wettstein-Knowles, P. (1991) *Proc. Natl. Acad. Sci. U. S. A.* **88,** 4114–4118
10. Aizo, M., Otoko, S., Miki, T., Soichiro, M., Junichi, M., and Minoru, O. (1992) Japanese Patent Application JP1992058889-A1
11. Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J., and Rutter, W. J. (1979) *Biochemistry* **18,** 5294–5299
12. Aviv, H., and Leder, P. (1972) *Proc. Natl. Acad. Sci. U. S. A.* **69,** 1408–1412
13. Kofod, L. V., Kauppinen, S., Christgau, S., Andersen, L. N., Heldt-Hansen, H. P., Dörreich, K., and Dalbøge, H. (1994) *J. Biol. Chem.* **269,** 29182–29189
14. Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual,* 2nd Ed., Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
15. Feinberg, A. P., and Vogelstein, B. (1983) *Anal. Biochem.* **132,** 6–13
16. Berka, R. M., Schneider, P., Golightly, E. J., Brown, S. H., Madden, M., Brown, K. M., Halkier, T., Mondorf, K., and Xu, F. (1997) *Appl. Environ. Microbiol.* **63,** 3151–3157
17. Berka, R. M., Rey, M. W., Brown, K. M., Byun, T., and Klotz, A. V. (1998) *Appl. Environ. Microbiol.* **64,** 4423–4427
18. Davis, R. W., Botstein, D., and Roth, J. R. (1980) *Advanced Bacterial Genetics, a Manual for Genetic Engineering,* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
19. D'Alessio, J. M., Bebee, R., Hartley, J. L., Noon, M. C., and Polayes, D. (Life Technologies Inc.) (1992) *Focus®* **14,** 76
20. Messing, J., Crea, R., and Seeburg, P. H. (1981) *Nucleic Acids Res.* **9,** 309–321
21. Giesecke, H., Obermaier, B., Domedy, H., and Neubert, W. J. (1992) *J. Virol. Methods* **38,** 47–60
22. Boel, E. Hjort, I., Svennson, B., Norris, F., Norris, K. E., and Fiil, N. P. (1984) *EMBO J.* **3,** 1097–1102
23. Ho., S. N., Hunt, H. D., Horton, R. M., Pullen, J. K., and Pease, L. R. (1989) *Gene (Amst.)* **77,** 51–59
24. Christensen, T., Wöldike, H., Boel, E., Mortensen, S. B., Hjortshøj, K., Thim, L., and Hansen, M. T. (1988) *Bio/Technology* **6,** 1419–1422
25. Straubinger, B., Straubinger, E., Wirsel, S., Turgeon G., and Yoder, O. (1992) *Fungal Genet. Newslett.* **39,** 82–83
26. Brody, H., Hansen, K., Lamsa, M. H., and Yaver, D. S. (1998) World Patent Application WO9811203
27. Bothwell, M. K., and Walker, L. P. (1995) *Bioresour. Technol.* **53,** 21–29
28. Frohman, M. A., Dush, M. K., and Martin, G. R. (1988) *Proc. Natl. Acad. Sci. U. S. A.* **85,** 8998–9002
29. Gurr, S. J., Unkles, S. E., and Kinghorn, J. R. (1987) in *Gene Structure in Eukaryotic Microbes* (Kinghorn, J. R., ed) pp. 93–139, IRL Press, Oxford
30. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402
31. von Heijne, G. (1986) *Nucleic Acids Res.* **14,** 4683–4690
32. Coutinho, P. M., and Reilly, P. J. (1994) *Protein Eng.* **7,** 393–400
33. Van Arsdell, J. N., Kwok, S., Schweickart, V. L., Ladner, M. B., Gelfand, D. H., and Innis, M. A. (1987) *Bio/Technology* **5,** 60–64
34. Henrissat, B. (1991) *Biochem. J.* **280,** 309–316
35. Henrissat, B., and Bairoch, A. (1993) *Biochem. J.* **293,** 781–788
36. Henrissat, B., and Bairoch, A. (1996) *Biochem. J.* **316,** 695–696
37. Gebler, J., Gilkes, N. R., Claeyssens, M., Wilson, D. B., Béguin, P., Wakarchuk, W. W., Kilburn, D. G., Miller, R. C., Jr., Warren, R. A. J., and Withers, S. G. (1992) *J. Biol. Chem.* **267,** 12559–12561
38. Davies, G., and Henrissat, B. (1995) *Structure* **3,** 853–859
39. Henrissat, B., and Davies, G. J. (1997) *Curr. Opin. Struct. Biol.* **7,** 637–644
40. Gilkes, N. R., Henrissat, B., Kilburn, D. G., Miller, R. C., and Warren, R. A. J. (1991) *Microbiol. Rev.* **55,** 303–315
41. Tomme, P., Warren, R. A. J., and Gilkes, N. R. (1995) *Adv. Microb. Physiol.* **37,** 1–81
42. Warren, R. A. J. (1996) *Annu. Rev. Microbiol.* **50,** 183–212
43. Hasegawa, S., and Nordin, J. H. (1969) *J. Biol. Chem.* **244,** 5460–5470
44. Matsuda, S., Kawanami, Y., Takeda, H., Ooi, T., and Kinoshita, S. (1997) *J. Ferment. Bioeng.* **83,** 593–595
45. Takehara, T., Inoue, M., Morioka, T., and Yokogawa, K. (1981) *J. Bacteriol.* **145,** 729–735
46. Oppermann, R. V. (1979) *Scand. J. Dent. Res.* **87,** 302–308

Curr Genet (2000) 38: 291–298

ORIGINAL PAPER

Susan L. Elrod · Aubrey Jones · Randy M. Berka
Joel R. Cherry

Cloning of the *Aspergillus oryzae* 5-aminolevulinate synthase gene and its use as a selectable marker

Received: 16 March 2000 / Accepted: 18 July 2000

Abstract The *hemA* gene encoding 5-aminolevulinate synthase, the first enzyme in heme biosynthesis, was cloned from *Aspergillus oryzae* and evaluated as a selectable marker for the transformation of filamentous fungi. Deletion of the *hemA* gene resulted in a lethal phenotype that could be rescued either by the supplementation of culture media with 5-aminolevulinic acid (ALA) or by transformation with the wild-type *hemA* gene, but not by growth on rich media, nor by the addition of exogenous heme. Transformation of a *hemA* deletion strain with the *hemA* gene linked to a lipase expression cassette yielded ALA prototrophs expressing lipase. The *hemA* gene can therefore be used as a selectable marker for the transformation of *A. oryzae*.

Key words *hemA* · Heterologous gene expression · Heme biosynthesis · Recombinant enzyme

Introduction

Filamentous fungi are widely used as efficient hosts for protein production, secreting some proteins at up to 30 g/l (Ward 1989). Harnessing these hosts for heterologous protein synthesis requires the introduction of a transgene, together with a marker gene to allow selection. Selectable markers can be grouped into three broad classes based on provision of an ability to: (1) grow in the presence of an antimicrobial substance, (2) synthesize a required nutrient, or (3) utilize a unique nitrogen or carbon compound for growth (for reviews on transformation and selection systems in fungi, see Esser and Mohr 1986; Cullen and Berka 1987; Timberlake and Marshall 1989; Gwynne and Devchand 1992). While these systems function effectively for the initial selection of transformed DNA, they require the addition of an antibiotic or the use of a defined growth medium for continuous selection during large-scale fermentation. Consequently, an alternate selection system functioning in rich media without costly additives was sought. Since commercially useful filamentous fungi, such as *Aspergillus* and *Fusarium*, are obligate aerobes, enzymes involved in respiration were chosen as targets for the development of such a system.

Heme, a cofactor required for functional respiratory cytochromes, is synthesized in an eight-step pathway that begins with the condensation of glycine and succinyl-CoA to form 5-aminolevulinate (ALA). The gene encoding 5-aminolevulinate synthase (ALAS), the enzyme catalyzing this reaction, has been cloned from a number of organisms including humans (Borthwick et al. 1984; Bawden et al. 1987), *A. nidulans* (Bradshaw et al. 1993) and *Saccharomyces cerevisiae* (Arrese et al. 1983; Urban-Grimal et al. 1986). Deletion of the *HEM1* gene encoding ALAS in *S. cerevisiae* is lethal, but cells can be rescued by the addition of exogenous ALA or heme, or by reintroduction of the wild-type gene (Urban-Grimal and Labbe-Bois 1981; Arrese et al. 1983; Bard and Ingolia 1984; Keng et al. 1986; Volland and Urban-Grimal 1988). *HEM1* has previously been used as a selectable marker in large-scale yeast fermentations, where oxygen limitation helped to maintain plasmid selection (Bard and Ingolia 1990).

Here we describe the cloning of the gene encoding the first enzymatic step in heme biosynthesis and its evaluation as a selectable marker in *A. oryzae*.

Communicated by B. G. Turgeon

S. L. Elrod (✉)[1] · A. Jones · R. M. Berka · J. R. Cherry
Novo Nordisk Biotech Inc.,
1445 Drew Avenue,
Davis, CA 95616, USA
e-mail: selrod@calpoly.edu

Present address:
[1]Biological Sciences Department, California Polytechnic State University, San Luis Obispo, CA 93407, USA

Materials and methods

Isolation of the *A. oryzae hemA* gene

The *A. oryzae hemA* gene was cloned by screening an *A. oryzae* genomic library with a probe generated by PCR amplification of the *A. nidulans hemA* gene. The genomic library was constructed by cloning *Tsp*509 I-cut DNA fragments (4-kb to 7-kb) into the λ-ZipLox cloning vector (Life Technologies, Bethesda, Md.). The probe was prepared by PCR amplification of *A. nidulans hemA* DNA (*A. nidulans* strain FGSC26) under conditions recommended in the DIG high prime DNA labeling and detection starter kit II (Boehringer Mannheim, Indianapolis, Ind.) using oligonucleotides ALAS3d 5′-TTTATGATGGAGGCCCTTCTCCAGCAGTCTC-3′ and ALAS4e 5′-CTATGCATTTAAGCAGCAGCCGCGAC-TGG-3′. Bacteriophage DNA from 7×10^4 plaques was transferred to duplicate circular nylon membranes (Nytran Plus, Schleicher & Schuell, Keene, N.H.) and hybridized with a digoxigenin (DIG)-labeled *A. nidulans hemA* probe. Membranes were hybridized at 42 °C in 5 × SSC, 0.1% Sarkosyl, 0.02% SDS, 1% Genius blocking agent (Boehringer Mannheim, Indianapolis, Ind.) and 30% formamide. Membranes were washed at room temperature twice in 5 × standard saline citrate (SSC) with 0.1% sodium dodecyl sulfate (SDS), followed by two washes in 2 × SSC, 0.1% SDS. Five clones were identified and excised into pZL derivatives according to the manufacturer's instructions (Life Technologies, Bethesda, Md.). These clones were found to overlap and span a 4.2-kb region containing the *hemA* gene (GenBank accession number AF152374). Sequencing was performed using a Perkin Elmer 377 automated sequencer using Big Dye chemistry (Perkin Elmer, Foster City, Calif.).

Creation of *hemA* deletion plasmid pSE52

Plasmid pSE17, a pZL derivative containing 3175 bp of *hemA* genomic insert [523 bp of upstream sequence, the *hemA* open reading frame (ORF) of 1911 bp and 722 bp of downstream sequence (Fig. 1)] was the starting point for the construction of a *hemA* deletion plasmid. A 4.1-kb *A. oryzae pyrG* fragment (including the *pyrG* ORF) was inserted into the *Eco*RI site of *hemA* in pSE17, positioning *pyrG* in the 3′ flanking region of *hemA*. The resulting plasmid was restricted with *Eco*47III and *Mlu*NI and re-ligated to create the *hemA*Δ::*pyrG* allele. The deletion allele was then PCR-amplified using the Expand PCR kit (Boehringer Mannheim, Indianapolis, Ind.) with primers SE48up 5′-AATGGTCAAAACTGGCTCCTAC-3′ and SE48dwn 5′-TGTACCTGTTCTTGGGCTGTC-3′ and subcloned into pCR2.1 (Invitrogen, Carlsbad, Calif.) to create pSE52 (Fig. 2). A linear 6.3-kb fragment containing the *hemA*Δ::*pyrG* allele could then be easily generated by restriction with *Sac*I and *Not*I.

Creation of *hemA*-lipase expression plasmid pSE54

Plasmid pSE54 was constructed by PCR amplification of *hemA* with primers SMup1 5′-GCTCTAGATACCTGTTCTTGGGC-TGTGAC-3′ and SM+term 5′-GCTCTAGATGGCCCCTT-CTATTGTTATTA-3′ from pSE17. The amplified product had *Xba*I sites at either end (underlined in the primer sequence), the *hemA* gene, and 491 bp of the native promoter. This *Xba*I fragment was inserted into pBANe8, a fungal expression plasmid containing the *Thermomyces languinosus* lipase gene (Boel and Huge-Jensen 1989) under the control of the Taka-amylase A gene promoter (Tada et al. 1991) to create plasmid pSE54 (Fig. 2).

Media and reagents

Minimal medium contained (per liter): 6 g $NaNO_3$, 0.52 g KCl, 1.52 g KH_2PO_4, 1 ml Cove trace elements (0.04 g $Na_2B_4O_7$-$10H_2O$,

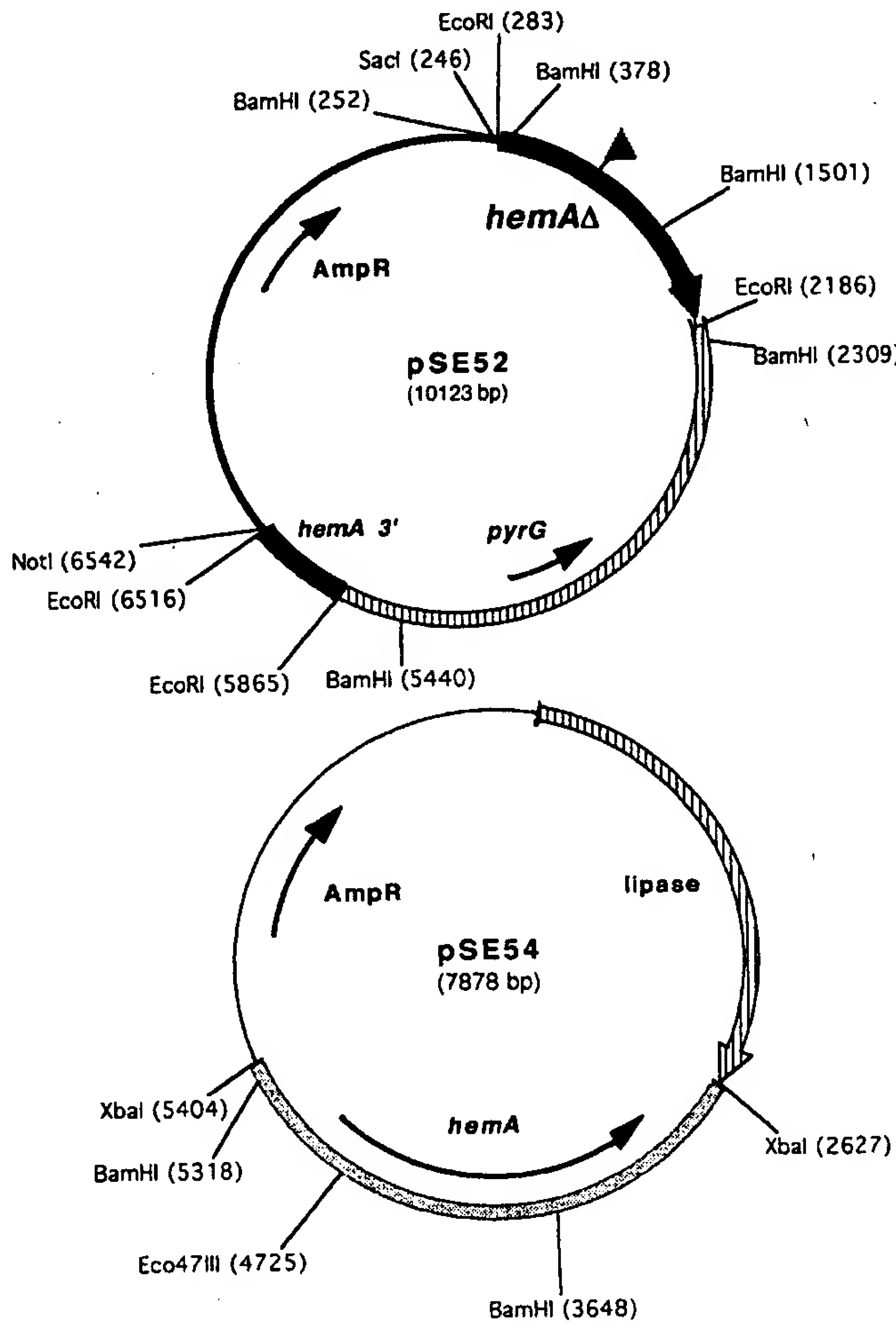


Fig. 2 Plasmid maps of the *hemA*::*pyrG* deletion plasmid, pSE52, and the *hemA*-lipase expression plasmid. Positions of relevant genes are indicated, ORFs are shown by *arrows* and the *closed triangle* indicates the location of the *hemA* deletion in pSE52. Nucleotide positions of relevant restriction sites are shown in parentheses

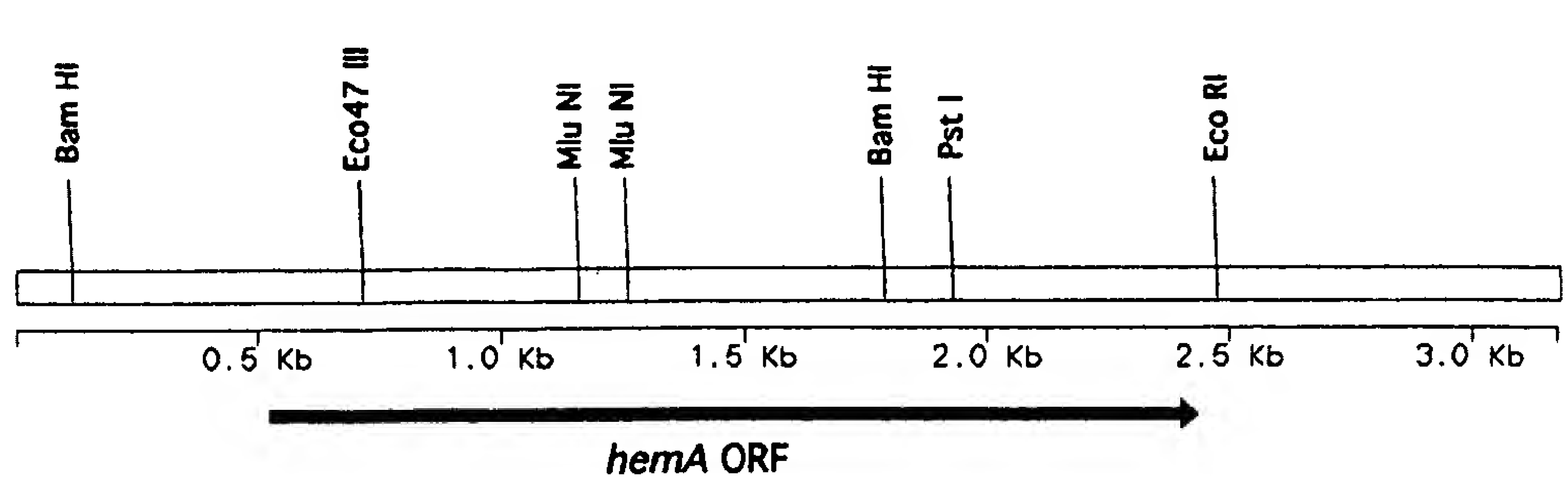


Fig. 1 Restriction map of the *Aspergillus oryzae hemA* gene *hemA*::*pyrG* deletion allele, showing the *hemA* genomic region in plasmid pSE17. The *arrow* indicates the putative *hemA* open reading frame (*hemA* ORF)

0.4 g $CuSO_4$-$5H_2O$, 1.2 g $FeSO_4$-$7H_2O$, 0.7 g $MnSO_4$, 0.8 g Na_2-MoO_2-$2H_2O$ and 10 g $ZnSO_4$-$7H_2O$ per liter), 25 g Noble agar, 1 M sucrose, 20 ml 50% glucose and 2.5 ml 20% $MgSO_4$, adjusted to pH 6.5. YEG contained (per liter): 5 g yeast extract, 20 g glucose and 1 M sucrose. MY25 medium contained (per liter): 1 × MY salts [2 g $MgSO_4$-$7H_2O$, 2 g K_2PO_4, 10 g KH_2PO_4, 2 g citric acid, 0.5 ml AMG trace elements (13.9 g $FeSO_4$-$7H_2O$, 8.5 g $MnSO_4$-H_2O, 14.28 g $ZnSO_4$-7 H_2O, 1.63 g $CuSO_4$, 0.24 g $NiCl_2$-6 H_2O and 3.0 g citric acid per liter) and 1 ml 10% $CaCl_2$-$2H_2O$], 1% yeast extract, 2.5% maltose and 0.2% urea, at pH 6.5. Quarter-strength MY25 contained 1 × MY salts and 0.25% of the yeast extract, maltose and urea concentrations from the above.

Hemin (Sigma Chemical Co., St. Louis, Mo.) and ALA (Porphyrin Products, Logan, Utah) were prepared immediately prior to use as stock solutions in 50 mM NaOH and water, respectively.

Fungal strains and transformations

Fungal transformations were performed as previously described (Christiansen et al. 1988) with the following exceptions: Novozyme 234 (Novo Nordisk, Bagsvaerd, Denmark) was used in a final concentration of 10 mg/ml during protoplasting without BSA, and 1 ml of SPTC [0.8 M sorbitol, 40% polyethylene glycol 4000 (BDH), 50 mM Tris-HCl and 50 mM $CaCl_2$, pH 8.0] was used instead of STC. Transformations of *hemA* deletion strain SE29-70 were performed using 10 μg of the indicated circular plasmid DNA (pSE17 or pSE54), and the protoplasts were plated onto minimal medium for growth at 34 °C.

The *hemA* deletion strain was created by transformation of PyrG$^-$ *A. oryzae* strain HowB425 (Brody et al. 1998) with a *hemAΔ::pyrG* deletion allele. The linear 6.3-kb *SacI*-*NotI* *hemAD::pyrG* fragment from plasmid pSE52 was isolated by agarose gel electrophoresis and used in the transformation reaction. Transformants were selected on minimal medium containing either 2.5 mM ALA or 5 mM ALA. ALA auxotrophy was determined by assessing growth of primary transformants on minimal medium lacking ALA at 34 °C. Strains showing marginal growth under these conditions were streaked for single colony isolation on minimal medium supplemented with ALA. Several strains were confirmed as *hemA* deletants by Southern analysis and one strain, SE29-70, was used for all subsequent transformations.

Southern hybridization analysis

Genomic DNA isolations

Genomic DNA was isolated as previously described (Wahleithner et al. 1996). For all Southern blots, 10 μg of restricted DNA was electrophoresed and transferred either to Nytran Plus (Schleicher & Schuell, Keene, N.H.) or to Hybond N (Amersham, Piscataway, N.J.) using 0.4 N NaOH in a TurboBlot apparatus, according to the manufacturer's instructions (Schleicher & Schuell, Keene, N.H.). Nylon membranes were rinsed in 2 × SSC after transfer, air-dried and UV cross-linked before hybridization with the indicated probes.

Probes

The *hemA* probe was generated by PCR amplification from pSE17 using primers hemAdelup1 5′-AGGCCTCTTGGGTTATG-AATG-3′ and hemAdeldwn1 5′-TGACCTGGAGATTAGACA-TAG-3′. This generated a 502-bp probe positioned between the *Bam*HI and *Eco*47III restriction sites in the *hemA* sequence (see Fig. 1). This probe was either labeled with DIG-labeled dUTP in a PCR reaction performed using a DIG-label PCR kit (Boehringer Mannheim, Indianapolis, Ind.) or with (α-^{32}P)-dCTP by random priming using the Prime-It II kit (Stratagene, La Jolla, Calif.). The radioactively-labeled probe was purified using a G50 midi column (5′-3′ Boulder, Col.). PCR products were purified using either QiaQuick Spin Columns (Qiagen, Valencia, Calif.) or GenElute columns (Supelco, Bellefonte, Pa.) and denatured prior to use.

Hybridizations

Membranes were prehybridized at 42 °C for 6 h in 15 ml of hybridization solution [5 × SSC, 0.1% sarkosyl, 0.02% SDS, 1% Genius blocking agent (Boehringer Mannhiem, Indianapolis, Ind.) and 50% formamide]. Denatured probe [1 ng DIG-labeled probe/ml hybridization solution or 1 × 10^7 cpm of (^{32}P)-labeled probe] was added to 10–15 ml of hybridization solution. Hybridizations were carried out for 16 h at 42 °C. The blot was washed twice at room temperature for 5 min each time in 2 × SSC, 0.1% SDS, then washed twice at room temperature for 5 min each time in 0.2 × SSC. 0.1% SDS and finally washed twice at 68 °C for 15 min each time in 0.1 × SSC, 0.1% SDS. The washed membranes were rinsed in 2 × SSC and were then exposed to Kodak X-OMAT AR film, followed by development using a Konica QX-70 automatic film processor.

Heterologous expression of lipase

Colonies growing in the absence of ALA on primary selection plates were used to inoculate 1 ml of quarter-strength MY25 medium in 24-well microtiter plates. Plates were incubated at 34 °C, shaking at 200 rpm, and culture broth samples were removed on days 4 and 7 for lipase enzyme assays. Expression was further evaluated using 25 ml of MY25 medium in 125-ml plastic flasks shaken at 200 rpm in a 34 °C room for 4 days.

Lipase enzyme assays

Assays were performed in microtiter plates, following dilution of the samples using 0.02% alpha olefin sulfonate and 100 mM Mops at pH 7.5, following the hydrolysis of ρ-nitrophenyl butyrate substrate (1.3 mM ρ-nitrophenyl butyrate in dimethyl sulfoxide, 4 mM $CaCl_2$-$2H_2O$ and 100 mM Mops, at pH 7.5) at 405 nm. Relative activity was calculated based on comparisons to known purified lipase stock solutions.

Results

Cloning of the *hemA* gene

The *A. oryzae hemA* gene was cloned from a genomic library that was screened using the *A. nidulans hemA* gene (Bradshaw et al. 1993) as a probe. Sequence analysis of several overlapping clones revealed a single contiguous ORF of 1911 nucleotides (Fig. 1). The coding sequence of *hemA* does not contain any introns, in contrast to the *A. nidulans hemA* gene which contains one intron near the 5′ end (Bradshaw et al. 1993). Southern blot analysis using low stringency hybridization conditions demonstrated it to be a single copy gene (data not shown). The 5′ untranslated sequence contains several pyrimidine-rich and AT-rich regions similar to other fungal genes (Gurr et al. 1987; Unkles 1992), as well as a CCAAT motif at position −249 (relative to the initiator ATG = +1). Also, an $(AC)_{35}$ repeat motif occurs in the 3′ untranslated region approximately 40 nucleotides after the stop codon. Similar repeats have been observed in subtelomeric, intron and promoter regions

of mammalian and *S. cerevisiae* genes but have no known function, although they have been implicated in gene amplification events (Passananti et al. 1987). Microsatellites such as these have also been reported as abundant sequences in the human, rat and tropical tree genomes (Condit and Hubbell 1991; Beckman and Weber 1992; Dib et al. 1996).

The nucleotide sequence of *hemA* encodes a predicted protein of 636 amino acids with a molecular weight of 68 kDa. Overall, the deduced amino acid sequence of the *A. oryzae* ALA synthase shares 81% identity with the *A. nidulans* homologue, 53% identity with the *S. cerevisiae* protein, and 37% identity with the human erythroid ALAS (Fig. 3). Sequences implicated in catalytic activity and pyridoxal phosphate co-factor binding (Ferreira and Gong 1995), characterized by an arginine- and glycine-rich loop (Fig. 3), are conserved. The first 35 amino acids of the *A. oryzae* protein are rich in serine, threonine, lysine and arginine, consistent with a function as a mitochondrial localization sequence (Fig. 3). Also, two potential heme regulatory motifs (HRMs) with the tri-peptide sequence CPV/F, occur in the predicted protein sequences from both the *A. nidulans* and *A. oryzae* genes; one in the presumed mitochondrial localization sequence and one in the beginning of the putative mature protein (Fig. 3).

Deletion of the *hemA* gene

A marked deletion allele, *hemA*Δ::*pyrG*, was constructed (see Materials and methods and Fig. 2) and used to

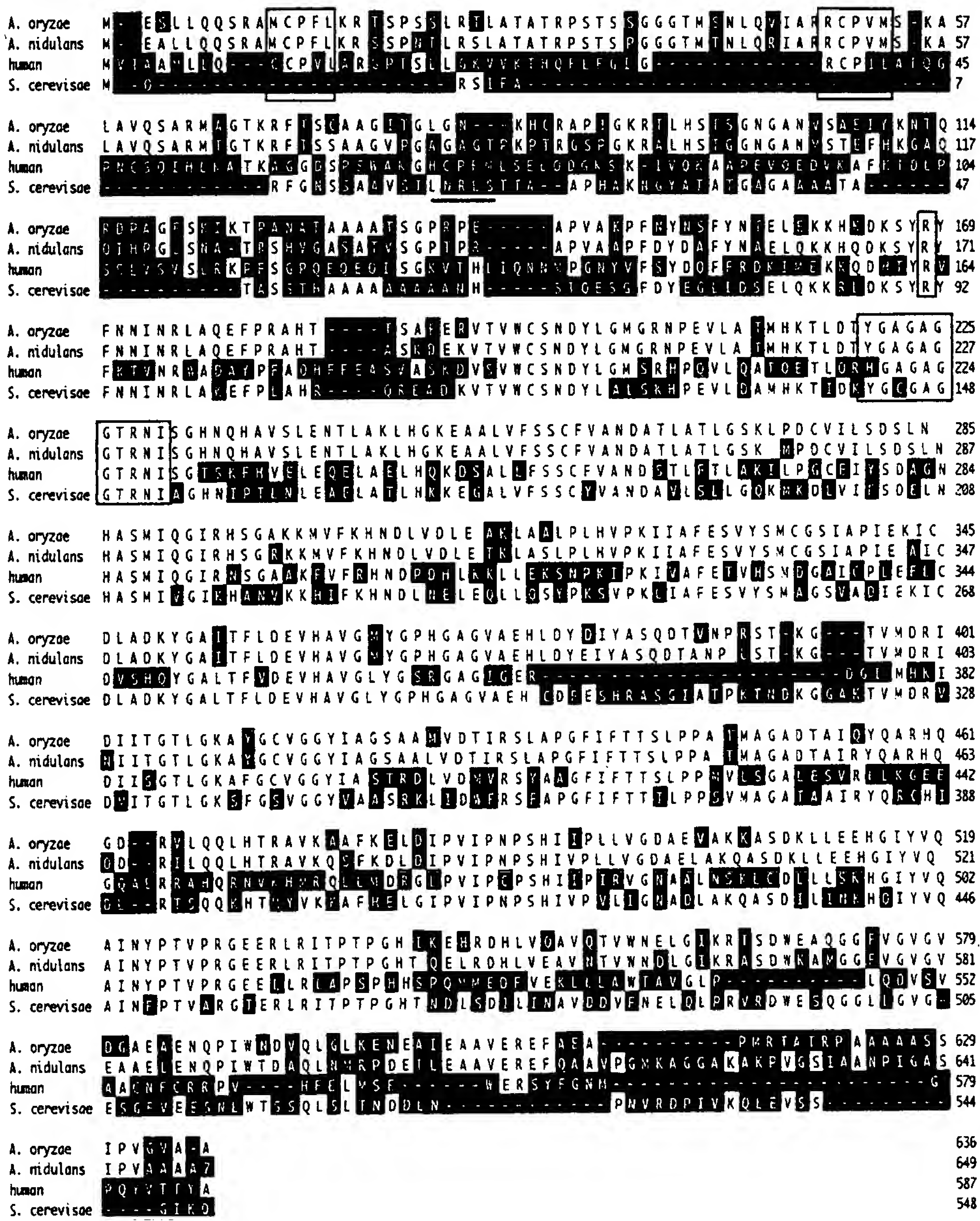


Fig. 3 *hemA* amino acid sequence alignment of 5-aminolevulinate synthase (ALAS) amino acid sequences from *A. oryzae*, *A. nidulans*, human erythroid (ALAS 2) and *Saccharomyces cerevisiae*. Differing amino acid residues are *shaded in black*. Conserved heme regulatory motifs (HRM; CPV/F), catalytic domain (conserved arginine) and glycine-rich loop (YGAGAGGTRNI) are *boxed*. An additional HRM sequence exists in the human sequence (*underlined*) which is not present in the fungal sequences. Sequences were aligned by the Clustal method with a gap penalty of 10, using MegAlign software (DNASTAR, Madison, Wis.)

transform a PyrG$^-$ *A. oryzae* strain. Transformants were plated onto minimal medium agar plates containing ALA to select for *pyrG* prototrophs while supporting growth of putative *hemA*-deleted strains. Screening of 240 primary transformants on minimal medium lacking ALA identified 11 strains with very poor growth on minimal medium. Two rounds of spore purification were required to generate strains that showed absolutely no growth on minimal medium. Southern hybridization of genomic DNA from five of these strains showed that one contained both a wild-type and a *hemA*Δ::*pyrG* allele, while four contained only the *hemA*Δ::*pyrG* allele (data not shown). Of these four, one strain (SE29-70) was chosen for all subsequent transformations. Confirmation that strain SE29-70 contains only the deletion allele is shown in Fig. 4, lane 1.

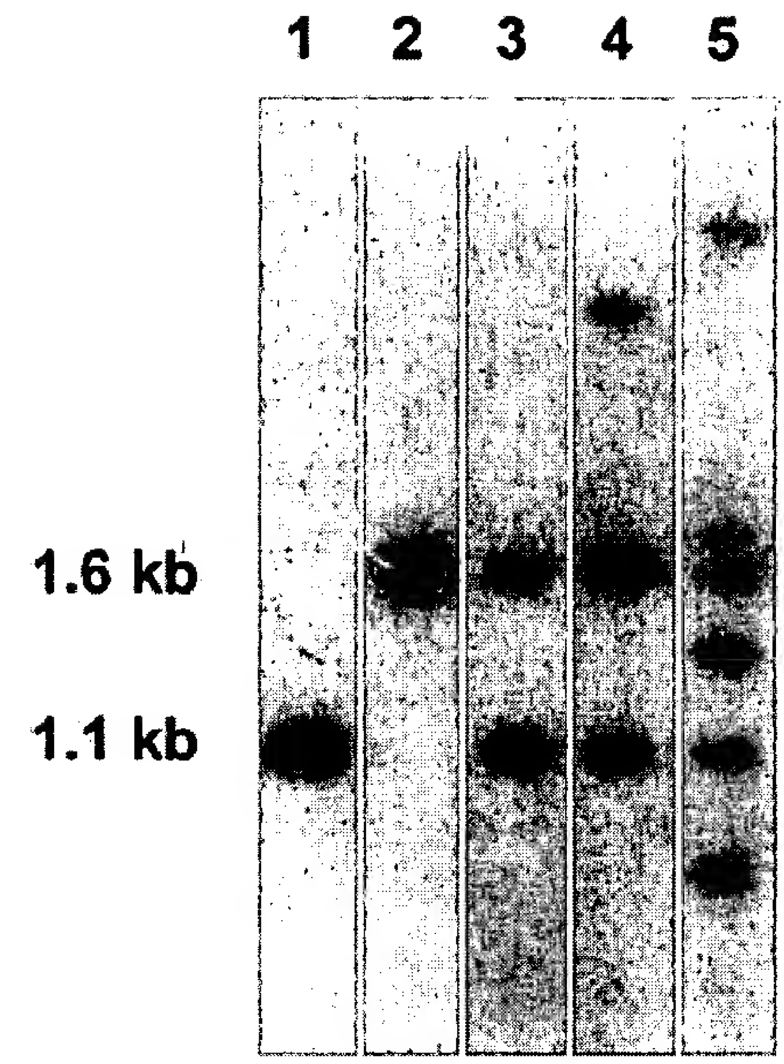


Fig. 4 Southern hybridization of *hemA*Δ::pyrG strains transformed with wild-type *hemA*. *Bam*HI digested genomic DNA was electrophoresed, transferred to a nylon membrane and hybridized with a *hemA* probe. *Lane 1 hemA* deletion strain (SE29-70), *lane 2* wild-type *A. oryzae*, *lanes 3–5* contain deletion strains transformed with plasmid pSE17. The wild-type *A. oryzae* strain shows the predicted 1.6-kb wild-type *hemA* band, while the *hemA* deletion strain shows a 1.1-kb band predicted for the deletion allele. All transformed strains contain both a wild-type (1.6 kb) and a deletion allele band (1.1 kb). *Lanes 4* and *5* also show extra bands, suggesting that the DNA has integrated into more than one site

Rescue of the lethal deletion phenotype

Concentrations of ALA as low as 30 μM (5.0 μg/ml) were sufficient to rescue the ALA auxotrophic phenotype when added to minimal medium agar. Hemin added to minimal medium in concentrations ranging from 0.05 mg/ml to 0.2 mg/ml had no inhibitory effect on the growth of wild-type strains, yet was unable to support growth of any deletion strain tested. Additional supplementation of minimal medium with 0.1 μg vitamin B_{12}/ml and/or 20 μg ergosterol/ml had no effect. The *hemA* deletion strains could not be maintained on rich medium such as MY25, indicating that media containing as much as 1% yeast extract do not contain sufficient ALA to rescue *hemA* auxotrophs. This suggests that nutrient-rich conditions used in commercial fermentations may be selective for the maintenance of *hemA*-linked expression vectors.

Complementation of *hemA* deletion by transformation with the *hemA* gene

Protoplasts of the *hemA*-deficient strain were transformed with plasmid pSE17 and plated on minimal medium lacking ALA. Transformed colonies were apparent after 2 days at 34 °C. Transformation efficiencies were similar to those observed using other fungal selectable markers, with a range of 5–30 transformants/μg DNA in independent transformations with different protoplast preparations. Control transformations with no DNA showed virtually no background colony formation. Most transformants contained a single copy of the *hemA* gene as shown in Fig. 4 (lane 3). However, one-third of transformed strains showed 1–3 extra copies of the *hemA* gene (Fig. 4, lanes 4, 5). All transformants tested were ALA prototrophs, as demonstrated by their ability to grow on minimal medium without ALA. These data suggest a correlation between the restoration of ALA prototrophy and the presence of at least one copy of the wild-type *hemA* gene.

The *hemA* gene as a selectable marker

In order to test the usefulness of *hemA* as a marker gene for fungal transformations, a plasmid was constructed that contained *hemA* and a heterologous lipase gene (pSE54; Fig. 2). The *hemA* gene used in this construct contained 523 bp of upstream DNA sequence. Use of this plasmid to transform *hemA* deletion strain, SE29-70, yielded 8.4 transformants/μg DNA. Eighty-eight percent (21/24) of primary transformants tested in 24-well plates produced levels of lipase ranging over 20–330 lipase units (LU)/ml (Fig. 5A). In shake-flask tests, the highest level of lipase produced by a *hemA*/lipase transformant was 1000 LU/ml. Previous work has demonstrated that untransformed *A. oryzae* strains do not produce any lipolytic activity (Huge-Jensen et al. 1989). In addition, levels produced by untransformed strains in the assay used in this study are routinely below the sensitivity limit, which is <3 LU/ml (Michael Lamsa, personal communication). Southern blot analysis of *hemA*/lipase transformants confirmed that they contained at least one copy of the wild-type *hemA* gene (data not shown). In contrast, a transformation efficiency of 3.8 transformants/μg DNA was obtained by transformation of strain SE29-70 with a plasmid containing the *amdS* gene and the same lipase expression cassette (pBANe8). Sixty-three percent (15/24) of these primary transformants tested in 24-well plates produced levels of lipase

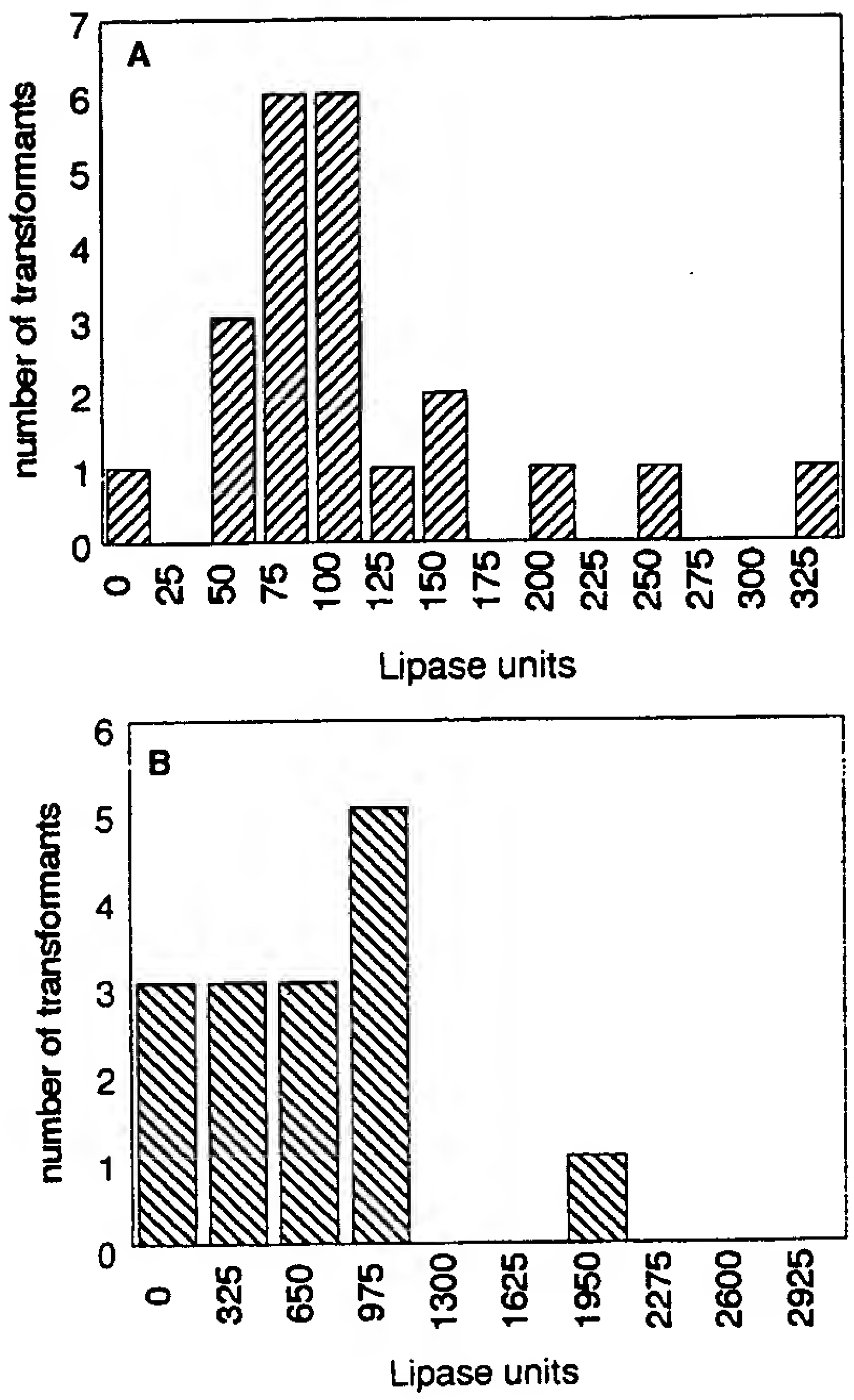


Fig. 5 A,B Lipase expression by *hemA* and *amdS* transformants. Histograms show the number of transformants from *hemA* transformation (A) or *amdS* transformation (B) producing various levels of lipase. Spores from isolated transformants were grown in 24-well plates; and samples removed on day 7 were tested for lipase activity

ranging over 95–2200 LU/ml (Fig. 5B). pBANe8 transformants produced up to 3700 LU/ml in shake-flasks.

Discussion

The goal of this work was to develop a new selectable marker system for use in filamentous fungi that did not require the use of antibiotics or specific carbon or nitrogen sources to maintain selection. Enzymes of the heme pathway were specifically targeted as possible selectable markers since reports from *S. cerevisiae* suggested that gene deletions in the heme biosynthetic pathway could be rescued by exogenous addition of pathway intermediates or by reintroduction of a pathway gene on a plasmid (Bard and Ingolia 1990). Furthermore, a selection system based on heme biosynthesis was presumed to link aerobic growth rate to transgene expression.

Cloning of the *hemA* gene encoding ALAS suggests that the mechanisms controlling expression of this gene are similar to those in other eukaryotes. A CCAAT motif found ca. 250 bp upstream of the start codon may be a site for the binding of the CCAAT-box transcription factor complex previously identified in *S. cerevisiae* (Keng and Guarente 1987), mammals (Guarente and Bermingham-McDonogh 1992) and plants (Edwards et al. 1998). Evidence for a similar transcription complex in *Aspergillus* (Kato et al. 1997) suggests that the CCAAT-box may have a similar function.

HRMs, such as those identified in the putative mitochondrial localization sequence, have been implicated in heme-mediated regulatory events at both the transcriptional and post-translational level. Heme has been shown to bind to related motifs in the *S. cerevisiae* HAP1 protein, resulting in dissociation of a repressor factor and activation of HAP1 in the presence of heme (Fytlovich et al. 1993; Zhang and Guarente 1994). Also, when present in leader sequences, HRMs have been shown to prevent import of ALAS proteins into mouse mitochondria via direct interactions with heme (Lathrop and Timko 1993; Zhang and Guarente 1994). This motif has further been postulated to bind heme in other proteins such as heme lyases (Steiner et al. 1996) and heme oxygenase-2 (McCoubrey et al. 1997). Interestingly, the yeast protein does not contain HRMs and only two motifs are present in *Aspergillus* proteins, while a third motif is found in mammalian sequences (Fig. 3; Ferreira and Gong 1995).

Deletion of the *hemA* gene proved to be lethal, and growth could only be restored by supplementing the growth media with ALA. Rich media containing yeast extract did not contain sufficient ALA to allow growth, suggesting that fermentation media would be selective for *hemA*-linked gene expression systems. Unlike *S. cerevisiae*, *hemA* deletion mutants could not be rescued by the addition of 0.2 mg exogenous heme/ml in *A. oryzae*, even when ergosterol and vitamin B_{12} were provided. In *S. cerevisiae HEM1* deletion strains can be rescued by 15 µg ALA/ml or 0.05 mg hemin/ml (Arrese et al. 1983; Volland and Urban-Grimal 1988). This difference may be attributed to differences between *S. cerevisiae* and filamentous fungal cell wall permeability or potential heme transport mechanisms.

It is interesting to note that initial attempts to delete the *hemA* gene failed to identify any primary transformants that were completely unable to grow in the absence of ALA. The inability to isolate "tight" auxotrophs immediately from primary transformants probably results from the transformation of multinucleate protoplasts and the subsequent segregation of the nuclei during serial culturing. After spore purification, strains carrying the *hemA* deletion were completely unable to grow in the absence of exogenous ALA, even when transferred directly from plates containing ALA.

Transformation of the *hemA* deletion strain with a plasmid containing the *hemA* gene and a lipase expression cassette demonstrated that *hemA* can be successfully used as a selectable marker. A plasmid containing both the lipase expression cassette and the *hemA* gene as a selectable marker produced primary transformants at an efficiency comparable to or better than control

transformations using the *amdS* selection system. The *hemA* transformants were easily distinguished and a high percentage (88%) of randomly chosen transformants produced detectable levels of lipase, indicating that they contained the transgene. This is in contrast to *amdS* transformants, which are often difficult to choose (63% were found to be transformed) due to the background of non-transformed colonies that can be prevalent, depending on the strain background. Lipase expression levels in independent transformants were, on average, 5- to 6-fold lower than those found using *amd*S as a selectable marker. These results are similar to those found by Christensen (1994) in a study comparing the *niaD* and *amdS* selectable markers, where *niaD* transformants produced lipase levels over a lower range than did *amdS* transformants. The *amdS* marker is known to produce higher copy number integrants than other markers, such as *argB* or *niaD* (Christensen 1994). In the current study, Southern analysis of *hemA*-lipase recombinants shows that the majority of *hemA* integrants are present in a single copy. This may be one reason for the lower levels of lipase observed in *hemA* transformants, when compared to *amdS* transformants. Although transformation with *hemA* can yield multiple integration events (as shown in Fig. 4), a single copy of *hemA* is apparently sufficient to adequately relieve the ALA auxotrophy. Manipulation of the *hemA* gene or promoter to create a debilitated allele may be useful in gaining recombinants with increased copy number, as has been demonstrated with the *LEU2* allele in *S. cerevisiae* (Erhart and Hollenberg 1983). Current efforts are focused on testing this hypothesis in *A. oryzae* and extending the use of *hemA* as a selectable marker in other fungal expression systems.

Acknowledgements The authors would like to thank Suzie Otani for generously providing the *A. oryzae* genomic library, Michael Rey and Elizabeth Golightly for DNA sequencing, Michael Lamsa and Rebecca Munds for lipase assays and Beth Nelson for supplying the lipase expression plasmid, pBANe8.

References

Arrese M, Carvajal E, Robison S, Sambunaris A, Panek A, Mattoon J (1983) Cloning of the 5-aminolevulinic acid synthase structural gene in yeast. Curr Genet 7: 175–183

Bard M, Ingolia TD (1984) Plasmid-mediated complementation of a 5-aminolevulinic-acid-requiring *Saccharomyces cerevisiae* mutant. Gene 28: 195–199

Bard M, Ingolia TD (1990) Novel DNA for expression of delta-aminolevulinic acid synthetase and related method. US Patent 4,902,620

Bawden MJ, Borthwick IA, Healy HM, Morris CP, May BK, Elliot WH (1987) Sequence of human 5-aminolevulinate synthase cDNA. Nucleic Acids Res 15: 8563

Beckman J, Weber J (1992) Survey of human and rat microsatellites. Genomics 12: 627–631

Boel E, Huge-Jensen B (1989) Recombinant *Humicula* sp. lipase. Publication 305,216. European Patent Office

Borthwick I, Srivastava G, Hobbs AA, Pirola BA, Brooker JD, May BK, Elliott WH (1984) Molecular cloning of hepatic 5-aminolevulinate synthase precursor. Eur J Biochem 144: 95–99

Bradshaw RE, Dixon SWC, Raitt DC, Pillar TM (1993) Isolation and nucleotide sequence of the 5-aminolevulinate synthase gene from *Aspergillus nidulans*. Curr Genet 23: 501–507

Brody H, Yaver D, Lamsa MH, Hansen K (1998) Cells having DNA insertion mutations which produce altered amounts of a polypeptide. PCT Patent Application WO 98/11203

Christensen T (1994) *Aspergillus oryzae* as a host for production of industrial enzymes. FEMS Symp 69: 251–259

Christensen T, Wøldike H, Boel E, Mortensen SB, Hjortshøj K, Thim L, Hansen MT (1988) High level expression of recombinant genes in *Aspergillus oryzae*. Biotechnology 6: 1419–1422

Condit R, Hubbell S (1991) Abundance of DNA sequence of two-base repeat regions in tropical tree genomes. Genome 34: 66–71

Cullen D, Berka RM (1987) Molecular genetics of commercially important filamentous fungi. Biotecnologie (Jan/Feb): 57–59

Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, et al (1996) A comprehensive genetic map of the human genome based on 5264 microsatellites. Nature 380: 152–154

Edwards D, Murray JAH, Smith AG (1998) Multiple genes encoding the conserved CCAAT-box transcription factor complex are expressed in *Arabidopsis*. Plant Physiol 117: 1015–1022

Erhart E, Hollenberg CP (1983) The presence of a defective *LEU2* gene on a 2 micron DNA recombinant plasmid of *Saccharomyces cerevisiae* is responsible for curing and high copy number. J Bacteriol 156: 625–635

Esser K, Mohr G (1986) Integrative transformation of filamentous fungi with respect to biotechnological application. Process Biochem (October): 153–159

Ferreira G, Gong J (1995) 5-aminolevulinate synthase and the first step of heme biosynthesis. J Bioenerg Biomembr 27: 151–159

Fytlovich S, Gervais M, Agrimonti C, Guiard B (1993) Evidence for an interaction between the CYP1(HAP1) activator and a cellular factor during heme-dependent transcriptional regulation in the yeast *Saccharomyces cerevisiae*. EMBO J 12: 1209–1218

Guarente L, Bermingham-McDonogh O (1992) Conservation and evolution of transcriptional mechanisms in eukaryotes. Trends Genet 8: 27–32

Gurr SJ, Unkles SE, Kinghorn JR (1987) The structure and organization of nuclear genes of filamentous fungi. In: Gene structure in eukaryotic microbes. (Special Publication, Society of General Microbiology) Oxford IRL Press, Oxford, pp 93–139

Gwynne D, Devchand M (1992) Expression of foreign proteins in the genus *Aspergillus*. In: Bennett J, Klitch M (eds) *Aspergillus*: biology and industrial applications. Butterworth-Heinemann, Stoneham, Mass

Huge-Jensen B, Andreason F, Christensen T, Christensen M, Thim L, Boel E (1989) *Rhizomucor miehei* triglyceride lipase is processed and secreted from transformed *Aspergillus oryzae*. Lipids 24: 781–785

Kato M, Aoyama A, Naruse F, Kobayashi T, Tsukagoshi N (1997) An *Aspergillus nidulans* nuclear protein, AnCP, involved in enhancement of Taka-amylase A gene expression, binds to the CCAAT-containing taaG2, amdS, and gatA promoters. Mol Gen Genet 254: 119–126

Keng T, Guarente L (1987) Constitutive expression of the yeast *HEM1* gene is actually a composite of activation and repression. Proc Natl Acad Sci USA 84: 9113–9117

Keng T, Alani E, Guarente L (1986) The nine amino-terminal residues of 5-aminolevulinate synthase direct B-galactosidase in the mitochondrial matrix. Mol Cell Biol 6: 355–364

Lathrop JT, Timko MP (1993) Regulation by heme of mitochondrial protein transport through a conserved amino acid motif. Science 259: 522–525

McCoubrey W, Huang T, Maines MD (1997) Heme oxygenase-2 is a hemoprotein and binds heme through heme regulatory motifs that are not involved in heme catalysis. J Biol Chem 272: 12568–12574

Passananti C, Davies B, Ford M, Fried M (1987) Structure of an inverted duplication formed as a first step in a gene amplification event: implications for a model of gene amplification. EMBO J 6: 16697–16703

Steiner H, Kispal G, Zollner A, Haid A, Neupert W, Lill R (1996) Heme binding to a conserved Cys-Pro-Val motif is crucial for the catalytic function of mitochondrial heme lyases. J Biol Chem 271: 32605–32611

Tada S, Gomi K, Kitamoto K, Kumagai C, Tamura G, Hara S (1991) Identification of the promoter region of the Taka-amylase A gene required for starch induction. Agric Biol Chem 55: 1939–1941

Timberlake WE, Marshall MA (1989) Genetic engineering of filamentous fungi. Science 244: 1313–1317

Unkles S (1992) Gene organization in industrial filamentous fungi. In: Kinghorn J, Turner G (eds) Applied molecular genetics of filamentous fungi. Blackie, Glasgow, pp 28–53

Urban-Grimal D, Labbe-Bois R (1981) Genetic and biochemical characterization of mutants of *Saccharomyces cerevisiae* blocked in six different steps of heme biosynthesis. Mol Gen Genet 183: 85–92

Urban-Grimal D, Volland C, Garnier T, Dehoux P, Labbe-Bois R (1986) The nucleotide sequence of the *HEM1* gene and evidence for a precursor form of the mitochondrial 5-aminolevulinate synthase in *Saccharomyces cerevisiae*. Eur J Biochem 156: 511–519

Volland C, Urban-Grimal D (1988) The presequence of yeast 5-aminolevulinate synthase is not required for targeting to mitochondria. J Biol Chem 263: 8294–8299

Wahleithner J, Xu F, Brown KM, Brown SH, Golightly EJ, Halkier T, Kauppinen S, Pederson A, Schneider P (1996) The identification and characterization of four laccases from the plant pathogenic fungus *Rhizoctonia solani*. Curr Genet 29: 395–403

Ward M (1989) Heterologous gene expression in *Aspergillus*. In: Nevalainen H, Penttila M (eds) Proc EMBO–ALKO workshop on molecular biology of filamentous fungi, vol 6. Foundation for Biotechnical and Industrial Fermentation Research, Helsinki, pp 119–128

Zhang L, Guarente L (1994) HAP1 is nuclear but is bound to a cellular factor in the absence of heme. J Biol Chem 269: 14643–14647

doi:10.1006/jmbi.2000.3550 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2000) 297, 233–249

JMB

Assessing Annotation Transfer for Genomics: Quantifying the Relations between Protein Sequence, Structure and Function through Traditional and Probabilistic Scores

Cyrus A. Wilson[1], Julia Kreychman[1] and Mark Gerstein[1,2]*

[1]*Department of Molecular Biophysics and Biochemistry*

[2]*Department of Computer Science, Yale University, 266 Whitney Avenue, PO Box 208114, New Haven, CT 06520, USA*

**Corresponding author*

Measuring in a quantitative, statistical sense the degree to which structural and functional information can be "transferred" between pairs of related protein sequences at various levels of similarity is an essential prerequisite for robust genome annotation. To this end, we performed pairwise sequence, structure and function comparisons on ~30,000 pairs of protein domains with known structure and function. Our domain pairs, which are constructed according to the SCOP fold classification, range in similarity from just sharing a fold, to being nearly identical. Our results show that traditional scores for sequence and structure similarity have the same basic exponential relationship as observed previously, with structural divergence, measured in RMS, being exponentially related to sequence divergence, measured in percent identity. However, as the scale of our survey is much larger than any previous investigations, our results have greater statistical weight and precision. We have been able to express the relationship of sequence and structure similarity using more "modern scores," such as Smith-Waterman alignment scores and probabilistic *P*-values for both sequence and structure comparison. These modern scores address some of the problems with traditional scores, such as determining a conserved core and correcting for length dependency; they enable us to phrase the sequence-structure relationship in more precise and accurate terms. We found that the basic exponential sequence-structure relationship is very general: the same essential relationship is found in the different secondary-structure classes and is evident in all the scoring schemes. To relate function to sequence and structure we assigned various levels of functional similarity to the domain pairs, based on a simple functional classification scheme. This scheme was constructed by combining and augmenting annotations in the enzyme and fly functional classifications and comparing subsets of these to the *Escherichia coli* and yeast classifications. We found sigmoidal relationships between similarity in function and sequence, with clear thresholds for different levels of functional conservation. For pairs of domains that share the same fold, precise function appears to be conserved down to ~40% sequence identity, whereas broad functional class is conserved to ~25%. Interestingly, percent identity is more effective at quantifying functional conservation than the more modern scores (e.g. *P*-values). Results of all the pairwise comparisons and our combined functional classification scheme for protein structures can be accessed from a web database at http://bioinfo.mbb.yale.edu/align

Keywords: bioinformatics; sequence similarity; percent identity; structure similarity; functional classification

Abbreviations used: EC, Enzyme Commission; EST, expressed sequence tags; SCOP, structural classification of proteins; GO, Gene Ontology Project.

E-mail address of the corresponding author: Mark.Gerstein@yale.edu

0022-2836/00/010233–17 $35.00/0

Introduction

The problem of genome annotation

Perhaps the most valuable information to be gained from a genome analysis is functional annotation of all the gene products. Unfortunately, of all the proteins whose sequences are known, functions have been experimentally determined for only a very small number (Andrade & Sander, 1997). Given the current size and accessibility of sequence and structure data, homologs of a newly sequenced gene's product can be identified *via* database searches, and probable structure and function assigned to the gene product (Bork *et al.*, 1998). This is based on the concept that sequence similarity implies structural and functional similarity. However, structural and functional annotations should be transferred with caution. If a protein is assigned an incorrect function in a database, the error could carry over to other proteins for which structure or function is inferred by homology to the errant protein (Brenner, 1999; Karp, 1996, 1998a). In large databases such an error can propagate out of control, presenting a serious quality control issue as we move to larger genomes from multicellular organisms.

Benchmarking fold and function recognition

Here, we used manually curated structural and functional classifications as standards in analyzing to what degree annotations of a protein's structure and function can be transferred to a similar sequence. The knowledge gained from the study can be used to establish confidence levels for structure and function prediction, improving our understanding of how long it will take to annotate accurately an entire genome.

Our simultaneous analysis of relationships between sequence and structure, sequence and function, and structure and function (Figure 1) may provide insight into paradigms for functional prediction other than that based alone on sequence similarity (Enright *et al.*, 1999).

Past results

Sequence-structure

The transfer of structural annotation is well characterized. Chothia & Lesk (1986, 1987) found that structural divergence, when expressed in terms of the RMS separation of matching alpha carbon atoms, was an exponential function of sequence divergence, expressed in terms of the fraction of residues that differed between sequences. The reliability of structural annotation transferred by homology, then, depends on the sequence identity of the homologous proteins (Chothia & Lesk, 1986). Flores *et al.* (1993), Russell & Barton (1994), and Russell *et al.* (1997) observed the same general trend, and also characterized the conservation of structural features other than the C^{α} backbone, such as secondary structure, accessibility and torsion angles. A paper by Wood & Pearson (1999) re-expressed the sequence-structure relationship in terms of statistically based "Z-scores" and found that this relationship had a simple linear form in terms of these scores. They also noted that protein families differed in detail in the slope of this linear relationship.

Others have focused on the limits of sequence comparison, specifically around the "twilight zone," the region of sequence similarity that does not reliably imply structural homology (Doolittle, 1987), and on establishing cut-offs for significant sequence similarity. Using the SCOP structural classification (Murzin *et al.*, 1995), Brenner *et al.* (1998) benchmarked the effectiveness of the popular FASTA and BLASTP programs and their probabilistic scoring schemes (i.e. the *e*-value) (Pearson & Lipman, 1988; Pearson, 1996; Altschul *et al.*, 1990, 1994; Karlin & Altschul, 1993). They found that in making fold assignments, the FASTA *e*-value closely tracked the number of false positives, i.e. the error rate, and that at a conservative *e*-value cut-off of 0.001, the FASTA program could detect nearly all the relationships that would be detected by a full Smith-Waterman comparison (Smith & Waterman, 1981). Specifically, they found that FASTA with a 0.001 threshold would find 16% more of the structural relationships in SCOP than would be found by standard sequence comparison with a 40% identity threshold. This rigorous benchmarking approach has been extended to assess transitive sequence comparison, through a third intermediate sequence and multiple-sequence matching programs such as PSI-blast (Park *et al.*, 1997, 1998; Gerstein, 1998a; Salamov *et al.*, 1999). In a related study Rost (1999) worked on characterizing the region after the twilight zone, which he called the "midnight zone". In a sense these benchmarking studies have culminated in the CASP fold recognition experiments (Moult *et al.*, 1997; Sternberg *et al.*, 1999).

Sequence-function

Although the exact dependence of functional similarity on sequence and structural similarity is not completely clear, initial indications of a gene product's function are most often based on simple sequence similarity (Bork *et al.* 1994, 1998). Often these are merely based on the best hit in database comparisons; see, for example, the annotation of some of the early genomes (Fraser *et al.*, 1995, 1998). However, possibilities for more robust annotation transfer are increasingly available. One looks at the pattern of hits amongst different phylogenetic groups (Tatusov *et al.*, 1997). Often these focus on the existence of key motifs and patterns associated with function (Zhang *et al.*, 1998; Bork & Koonin, 1996; Attwood *et al.*, 1999).

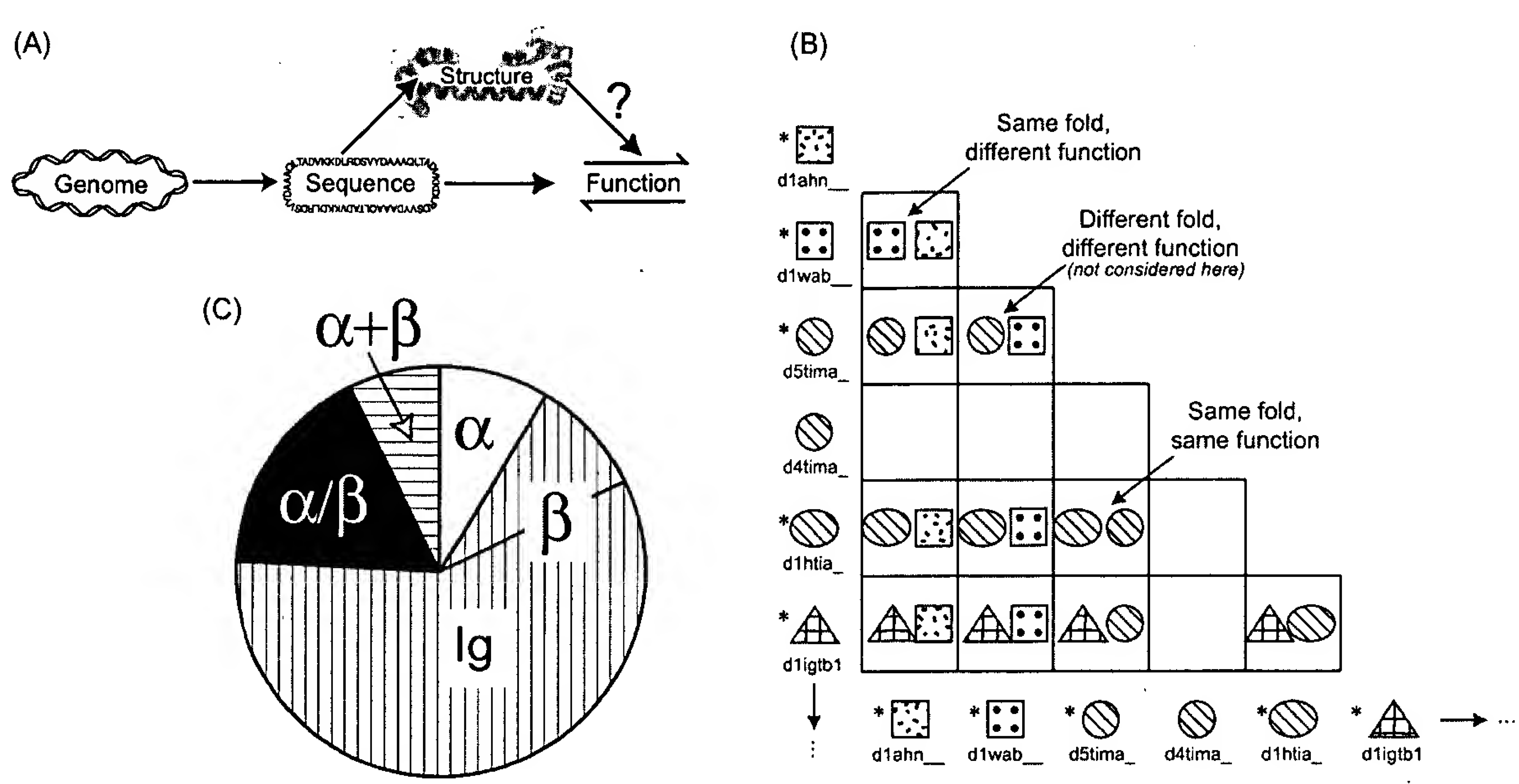


Figure 1. This Figure schematically depicts certain aspects of our comparison methodology. (a) The paradigm relating sequence to structure to function. There has not been as much assessment of functional annotation transfer based on structure as there has been with sequence-based structural and functional annotation transfer. (b) How we conceptualized our analysis in terms of pairs. A few examples of SCOP domains (identified on the left and bottom) are included from our comparison. In the Figure the shape represents fold, and the pattern represents function. We have highlighted some example categories of pairs: a pair that shares fold and function, a pair that shares fold but not function and a pair that shares neither fold nor function. The latter category of pairs is not considered in our investigation; we looked only at paired domains with the same fold. In constructing our pairs, we used only a representative set of SCOP domains. This is illustrated in the Figure by the domains flagged with asterisks. Note, in particular, that the SCOP domain d4tima_ is not paired with anything because it is represented by d5tima_, which is the same species and protein. For each level of pairs (fold, superfamily, family), cluster representatives were chosen for the level below: (i) for family pairs, one representative was selected from each species/protein, the level below, and then paired with all the other representatives within its family; (ii) for superfamily pairs, one representative was chosen from each family, unless there were domains in the family that shared less than 40 % sequence identity, in which case additional representatives were included, each not more than 40 % identical with the other representatives from the family (this occurs, for instance, for the globins); and (iii) likewise for fold pairs, one representative was chosen from each superfamily, more if there were domains with less than 40 % sequence identity. (c) Subdivides the pairs into the four SCOP classes from which they were composed: (i) all-α, domains consisting of α-helices; (ii) all-β, domains consisting of β-sheets; (iii) α/β, domains with integrated α-helices and β-strands; and (iv) α + β, domains with segregated α-helices and β-strands. We initially set apart the immunoglobulins from the rest of the all-β pairs because we realized that their large number biases our data. However, we compared the results for the immunoglobulin pairs to all other pairs and found that they generally exhibit the same behavior as the other pairs. Therefore we decided to leave them in the comparison.

Sequence-structure-function

One way that the better-defined sequence-structure relationship can assist in function prediction is initially to predict the structure of an uncharacterized sequence and then predict the function based on the limited repertoire of functions known to occur with that structure. To some degree this was achieved by Fetrow and co-workers (Fetrow *et al.*, 1998; Fetrow & Skolnick, 1998). They predicted structural profiles based on threading and *ab initio* methods, and then searched with these against profiles of known structures in order to predict function.

In related work, Russell *et al.* (1998) discussed using identification of structural binding sites in predicting protein function. In a comprehensive study, Hegyi & Gerstein (1999) investigated to what degree folds were associated with functions. They found that most folds were associated with one or two functions with the exception of a few special folds, such as the TIM barrel, that could carry out numerous functions. Furthermore, they found that particular folds were often confined to distinct phylogenetic groups, an additional fact that can feed into an integrated sequence-structure-function analysis (Gerstein & Hegyi, 1998; Gerstein, 1997, 1998b,c).

Here, we look at pairwise comparisons of protein sequence, structure and function among proteins that share the same fold. We assess the

trends relating sequence, structure and function and consider the implications for structural and functional annotation transfer.

New developments: probabilistic scoring and growth of the databank

The past studies regarding sequence, structure and function relationships often used RMS separation and percent sequence identity (or a linear variant of it, such as the fraction of mutated residues) to express similarities in structure and in sequence, respectively. However, it has become increasingly common to use probabilistic scoring schemes (P-values) to express the quality of a match in terms of statistical significance rather than an arbitrary raw score such as percent identity (Pearson, 1998; Karlin & Altschul, 1990, 1993; Karlin *et al.* 1991; Altschul *et al.* 1994; Bryant & Altschul, 1995; Abagyan & Batalyov, 1997). With P-values, scores from different investigations can be compared in a common framework. Recently, it was found that sequence and structure similarity significance can be expressed as P-values in the same unified statistical framework (Levitt & Gerstein, 1998). Here, we use such probabilistic scoring methods to overcome the limitations of the more traditional scores.

Another recent development is the tremendous growth in the number of solved structures. The RCSB Protein Data Bank (Bernstein *et al.* 1977) now contains more than 10,000 protein structures. These structures are broken into more than 18,000 domains, and then domains that share a fold are paired up with each other for comparison (Figure 1(b)). Here, we survey ~30,000 pairs of protein domains that are known to have the same fold, approximately 1000 times the number compared by Chothia & Lesk (1986). The large scale of this comparison affords greater statistical weight to the results.

Alignment of 30,000 pairs from SCOP

The basic unit of comparison: a pair of protein domains

The protein domains that we studied were classified by SCOP, a Structural Classification of Proteins (Murzin *et al.* 1995; Brenner *et al.* 1996; Hubbard *et al.* 1997), a hierarchy of five levels: (i) class, domains that have the same secondary structural content (all-α, all-β, α/β, or $\alpha + \beta$); (ii) fold, domains that geometrically share the same tertiary fold; (iii) superfamily, domains descended from the same ancestor (but which lack measurable sequence similarity); (iv) family, domains in the same protein sequence family (which have appreciable sequence similarity); and (v) species and protein.

Pairs of protein domains that are grouped together at the fold, superfamily or family level form the basic unit of our comparisons.

Selection of pairs

There is potentially a huge number of pairs of domains that can be constructed out of the relationships in SCOP. For instance, in the current version of SCOP there are ~3.9 million potential pairs between domains sharing the same fold. Most of these are between nearly identical structures. In order to keep the number of pairs manageable, we used a straightforward clustering scheme, described in the legend to Figure 1. We selected 29,454 representative pairs from the total in SCOP. To achieve a wide range of similarities, we constructed the pairs on three levels of the SCOP hierarchy: (i) family pairs, 19,542 pairs of domains in the same family; (ii) superfamily pairs, 4220 pairs of domains in the same superfamily but different families; and (iii) fold pairs, 5692 pairs of domains in the same fold but different superfamilies.

All the selected domains were at least 50 residues in length and were drawn from the four major SCOP secondary-structural classes: all-α, all-β, α/β, and $\alpha + \beta$ (Figure 1(c)).

We automatically aligned each of our selected domain pairs twice, once by global Needleman-Wunsch sequence comparison (Needleman & Wunsch, 1971; Myers & Miller, 1998) and then by structure (Gerstein & Levitt, 1996, 1998), calculating scores for sequence and structural similarity.

Web-accessible database

The results of all the pairwise comparisons are available *via* a searchable database on the web at http://bioinfo.mbb.yale.edu/align The query engine allows searches of individual SCOP pairs, all pairs that include a given SCOP domain, or all pairs containing any SCOP domain contained in a given PDB entry.

Traditional scores: RMS and percent identity

The sequence-structure relation, as expressed by the root-mean-square (RMS) of the aligned C^{α} distances and percent sequence identity, has been previously characterized as an exponential function by Chothia & Lesk (1986) and others (Flores *et al.* 1993; Russell & Barton, 1994; Russell *et al.* 1997). As Figure 2 illustrates, our data display a similar trend. (Exact equations are given in the legend to Figure 2.) However, we have one thousand times as many data points as in Chothia and Lesk's original study (30,000 as opposed to 30).

The main difference between our results and the previous studies is due to differences in RMS "trimming" methods. By trimming we refer to the process of removing the worst-fitting aligned atoms from the RMS calculation, to arrive at a structural "core." This was first developed in Lesk's sieve-fit procedure (Lesk & Chothia, 1984) and has been refined in numer-

ous studies (e.g. Gerstein & Altman (1995)). This is done because the small distances between well-matched alpha carbon atoms have much less of an effect on the RMS than do the very large distances between poorly matched atoms. The untrimmed score of divergent protein domains is then concerned primarily with the poorly matched residues instead of the conserved core. Trimming alleviates this effect by restricting the RMS calculation to include only those residues believed to be in the conserved core. However, the degree of trimming is to some extent arbitrary, and this choice affects the baseline of the reported RMS scores. Here we considered only the better half (50 %) of matched residues in a given pair of protein domains. Chothia & Lesk (1986) chose a somewhat different threshold. Figure 2(c) and (d) demonstrate the effect of trimming.

Analogous alignment similarity scores: Smith-Waterman score and structural comparison score

The dependence of the RMS separation on trimming method restricts its usefulness in comparing data. Likewise, there are many problems with using percent identity as a measure of sequence similarity. For instance, a match of non-identical but still similar residues (e.g. Arg *versus* Lys) scores the same as one between completely different residues (e.g. Arg *versus* Val), and gaps do not enter in the score calculation. Consequently, we now turn to alignment similarity scores, which eliminate some of the problems with traditional scores.

For sequence alignments, an alignment score is defined as the sum of the similarity matrix values for the alignment, minus the total gap penalty. This is sometimes called the Smith-Waterman score (Smith & Waterman, 1981). An analogous alignment score for structure is the structural comparison score, described by Levitt & Gerstein (1998). We will refer to these two similarity scores as S_{seq} and S_{str}, respectively. Note that they both increase for more similar pairs, whereas RMS increases for more divergent pairs. Specifically, S_{str} is the score maximized by the structural alignment program we used (Gerstein & Levitt, 1998). It can be calculated from any pair of aligned structures according to the function:

$$S_{str} = M\sum\left(\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2} - \frac{N_{gap}}{2}\right) \quad (1)$$

M and d_0 are constants, usually set to 10 and 5 Å, N_{gap} is the number of gaps in the alignment, d_i is the distance between each aligned pair of C^α atoms, and the sum is carried over all aligned pairs, i.

The main advantage of S_{str} over RMS in describing structural similarity is that the C^α to C^α distance, d_i, appears in the denominator of the calculation. This means that the smallest distances, corresponding to the best matches in the conserved core, are most significant in determining the score. Hence, the need for trimming is eliminated. S_{str} is also advantageous because it takes gaps into account and because of the fundamental analogy between this score and S_{seq}.

Figure 3(a) displays the relationship between structural and sequence similarity as expressed by S_{str} and S_{seq}. Figure 3(c) and (d) show calibration curves relating each of these scores back to approximate RMS separation and percent identity, respectively. Calibration curves help one get an intuitive feel for the degree of relationship in terms of the more traditional scores. Figure 3(b) adds a third axis, alignment length, and demonstrates that S_{str} depends greatly on this quantity. Although S_{str} and S_{seq} are "better" scores than RMS and percent sequence identity, the heavy dependence of both of these on length limits their usefulness in many situations. In other words, two pairs of similar domains with equal percent sequence identities but different lengths can have drastically different S_{seq} scores.

Probabilistic scores: *P*-values expressing the significance of sequence and structure similarity

Probabilistic scores can, to a great degree, overcome the length-dependence problems associated with the alignment scores. Probabilistic measures are advantageous because they express similarity not by an arbitrary "score" but by a statistical significance: the likelihood that such a similarity could be achieved by chance. This likelihood is also called the "*P*-value." We used calculations (described in detail in the legend to Figure 4) based on those given by Levitt & Gerstein (1998) to obtain *P*-values based directly on S_{str} and S_{seq}; we refer to these calculated *P*-values as P_{str} and P_{seq}, respectively. For P_{seq} we could equally well have used the numbers from one of the popular sequence search programs (i.e. BLAST or FASTA) as all these values have been shown to be perfectly proportional to each other (Levitt & Gerstein, 1998; Brenner *et al.* 1998).

P_{seq} and P_{str} can be used to express the relationship between structure and sequence similarity on a more fundamental level. Figure 4(a) shows a log-log (base 10) plot of P_{str} against P_{seq}. Because it is log-log, trends can be visualized as straight lines. Two straight lines are necessary to fit the points well, with the discontinuous boundary between the lines located at the beginning of the twilight zone. The different slope of the line at low sequence similarity reveals that in the twilight zone there is a different relationship between the significance of structural similarity and that of sequence similarity. In particular, for domain pairs

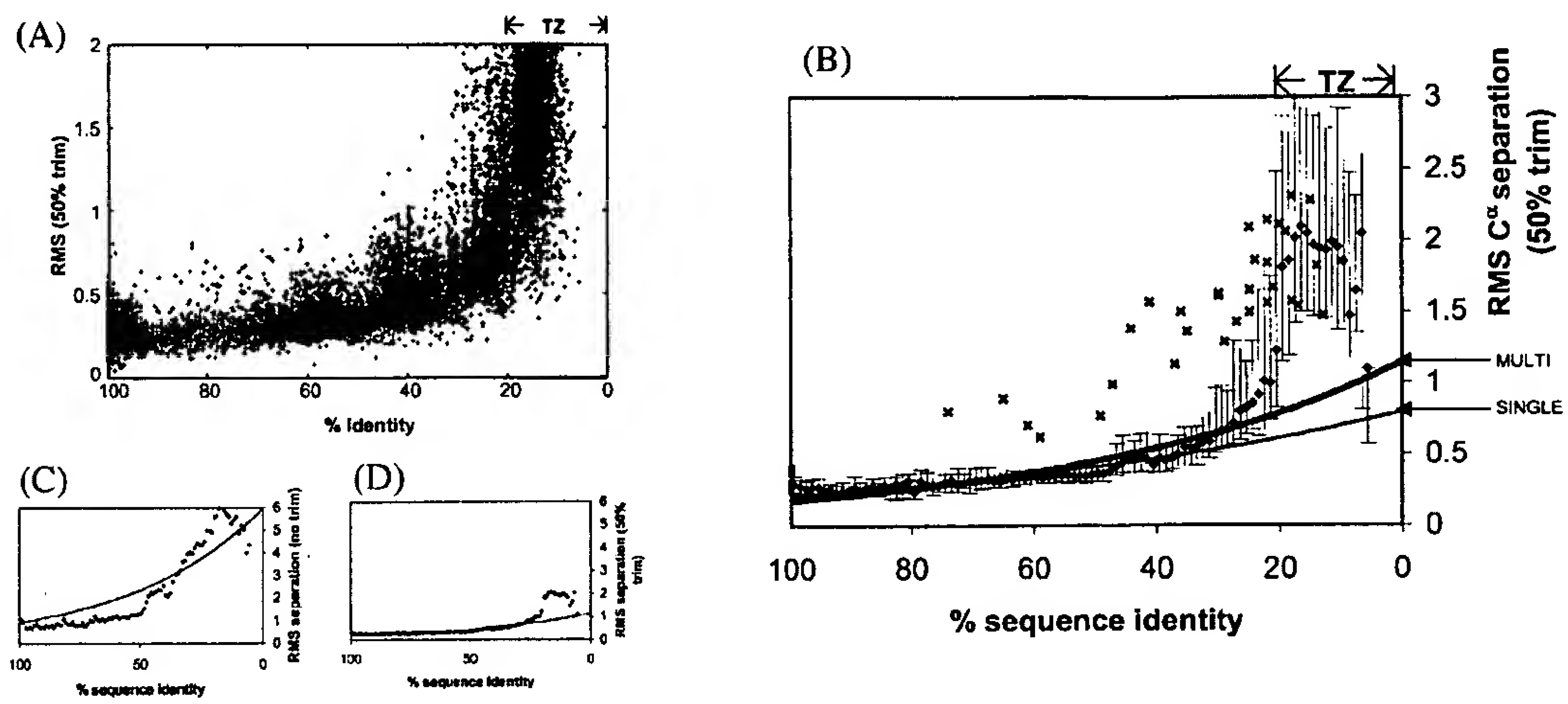


Figure 2. RMS as a function of percent identity. (a) A simple scatter plot of our pairs, relating RMS separation to percent sequence identity. This is similar to the presentation given by Chothia & Lesk (1986), but in this survey we looked at 30,000 pairs, 1000 times the number they compared. Outliers (pairs with RMS scores further than two standard deviations from the mean for their percent identity) are excluded from this graph; they represent domains that are very closely related with the exception of a conformational change. (b) A simplified graph with a number of fits to the data. For each percent identity bin we show the median RMS value, indicated by (◆) and the top and bottom quartile RMS values, indicated by the bars. Two fits are drawn through the median RMS values. The thin line, labeled SINGLE, is a simple exponential fit through the medians. It has the form:

$$R = 0.21e^{0.0132H}$$

where R is the RMS deviation after least-square fitting, H is the percent difference between the sequences (H for Hamming distance), and $H = 100\% - I$, where I is the percent sequence identity. The thick line, labeled MULTI, is a multigraph fit, which is described in the legend to Figure 4. The relation between RMS and percent identity according to this fit is expressed by the equation:

$$R = 0.18e^{0.0187H}$$

The twilight zone of sequence identity and below is labeled TZ. In this region, sequence similarity is not significant and not reliable for predicting structural similarity. This is why the median values in this area of the graph deviate significantly from the fits, which consider only data above 20 % sequence identity. For reference we include the original data points from Chothia and Lesk's, 1986 paper (A.M. Lesk, personal communication), indicated by X. Their data follow the form:

$$R = 0.40e^{0.0187H}$$

The difference between the Chothia & Lesk trend and our relationship is due to the different trimming methods used in calculating the RMS score. Chothia and Lesk imposed a 3 Å cut-off in determining the conserved core residues; we defined the core as the better matching (in terms of C^{α} distances) half (50 %) of the residue pairs. (c) and (d) The effect our trimming has on median RMS values. The RMS values in (c) are calculated from all the matched residues in each pair; the values in (d) are calculated from the better matching 50 % of the residues.

in the twilight zone (according to the percent identity to P_{seq} calibration in Figure 4(b)), structural similarity is more significant than sequence similarity (having a smaller P-value or more negative log P-value). In contrast, for pairs with more than ~30 % identity, the situation is reversed, with a given pair having more significant sequence similarity than structural similarity. One possible interpretation of this reversal is as follows. Structure is always more highly conserved than sequence, so usually a given amount of structural similarity is not as significant as a corresponding amount of sequence similarity. However, this is true only when meaningful sequence similarity actually exists; thus, it does not apply in the twilight zone, where sequence similarity is by definition not significant. Note that all pairs in our comparison share at least the same fold, implying that they always have a significant amount of structural similarity.

In other words, for closely related sequences, differences in sequence similarity are more meaningful, whereas for highly diverged sequences that share the same fold, the differences in structural similarity are more significant.

Fitting two lines to the P_{str} *versus* P_{seq} graph suggests that the same might be done for other scoring schemes. It is possible to some degree to fit

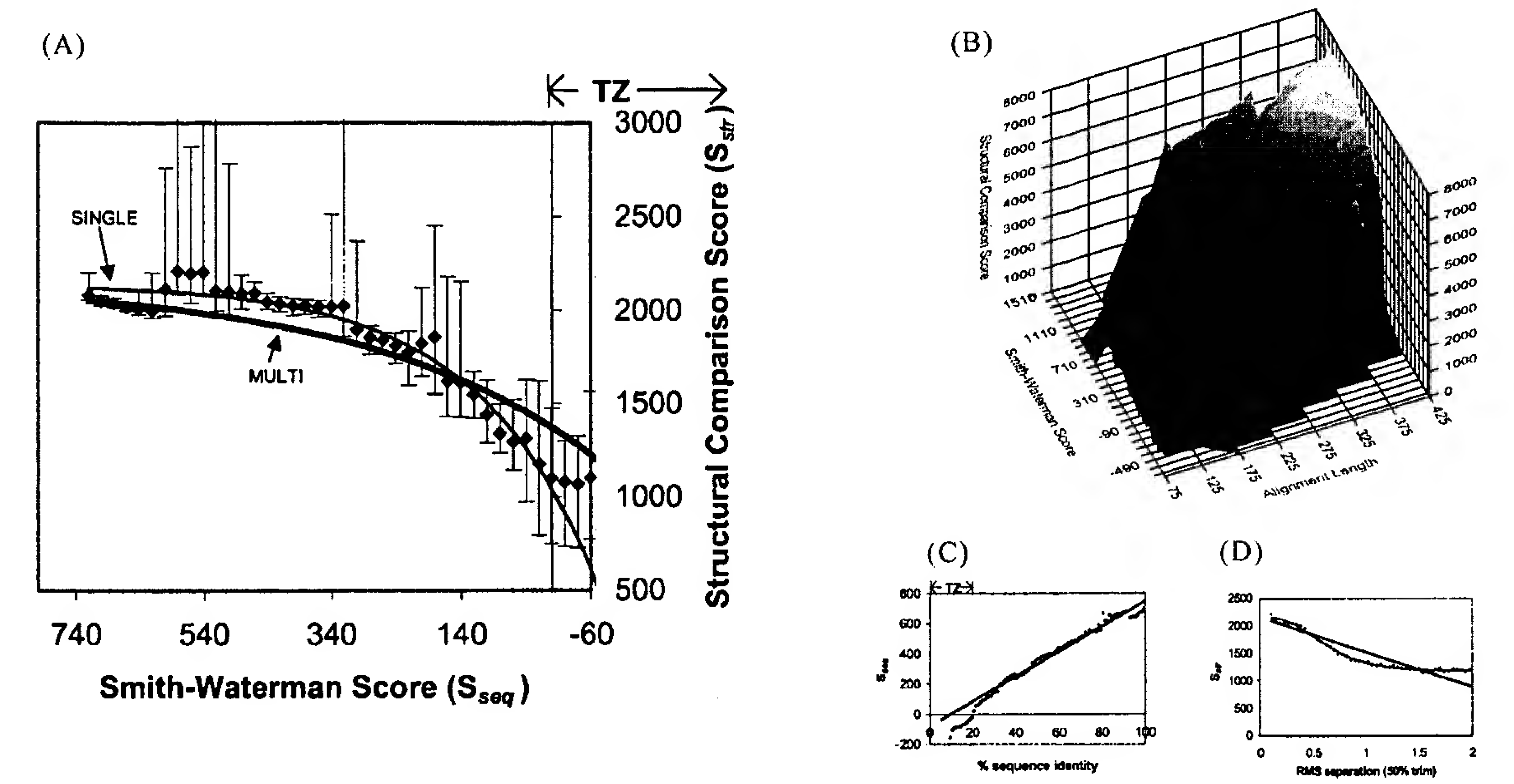


Figure 3. Similarity scores: structural comparison score as a function of Smith-Waterman score. Alignment similarity scores S_{str} and S_{seq} have certain advantages over RMS and percent identity scores for expressing the sequence-structure relation. S_{str} is calculated according to equation (1) in the text (Gerstein & Levitt, 1998; Levitt & Gerstein, 1998). S_{seq} is calculated using the BLOSUM50 matrix (Henikoff & Henikoff, 1992) with gap opening and extension penalties of −12 and −2, respectively. (a) This is analogous to (b) in Figure 2. From the original 30,000 pairs we show the median S_{str} value for each S_{seq} bin, along with quartile bars above and below. Again the twilight zone and below is labeled TZ. The thin line, marked SINGLE, is a simple fit to the median S_{str} values in this graph; it has the form:

$$S_{str} = 2144 - 1106\exp(-0.00544S_{seq})$$

The thick fit, marked MULTI, is the multigraph fit, explained below. It follows the equation:

$$S_{str} = 2157 - 787\exp(-0.0028S_{seq})$$

The equations presented here provide an approximation of the observed trends; as (b) illustrates, they are nothing more than simple approximations. The main disadvantage of S_{str} as a measure of structural similarity is its heavy length dependency for pairs of structurally similar protein domains. (b) Surface plot of the median S_{str} as a function of S_{seq} and alignment length (the number of matched residue pairs). It is clear that the size of the aligned domains plays a major role in the resulting S_{str}, even though our fits do not take length into account. (c) and (d) Relate S_{seq} and S_{str} to the more familiar percent identity and RMS measures. The fits were used to convert between scoring schemes in constructing the multigraph fit. We derived the multigraph fit in order to create one set of equations and parameters that would relate sequence and structural similarity using either the percent identity and RMS scheme or the S_{seq} and S_{str} scheme, and allow translation between them. We simultaneously performed least-squares fits to the median values in four graphs: Figures 2(b) and 3(a) and the calibrations of S_{seq} to percent identity and S_{str} to RMS, (c) and (d), respectively. In all cases, we ignored data in and below the sequence identity twilight zone (labeled TZ). The parameters in (a) are dependent on the parameters in Figure 2(b) *via* the mentioned calibrations.

the traditional RMS *versus* percent identity graph (Figure 2) with two straight lines instead of an exponential cruve. However, in this case, we opted for the more conventional presentation.

Class differences

The division of SCOP into classes based on secondary-structural composition allows easy investigation as to whether there are any deviations from the common similarity relationships on account of secondary-structure characteristics. Figure 5(a) reveals that secondary structural composition does not markedly affect the trends in sequence and structure similarities. This is consistent with the data given by Wood & Pearson (1999). However, the larger average length of α/β domains compared with domains in the other classes results in a deviation in the length-dependent S_{str} (Figure 5(b)). The consistency among length-independent scores applies for certain individual folds as well. The immunoglobulin fold makes up an appreciable fraction of all the β-pairs (Figure 1(c)), yet the results are not affected if these pairs are left out.

Linking sequence and structure to function

Difficulties of functional comparison

There is a clear, well-characterized relationship between sequence and structure similarity, which

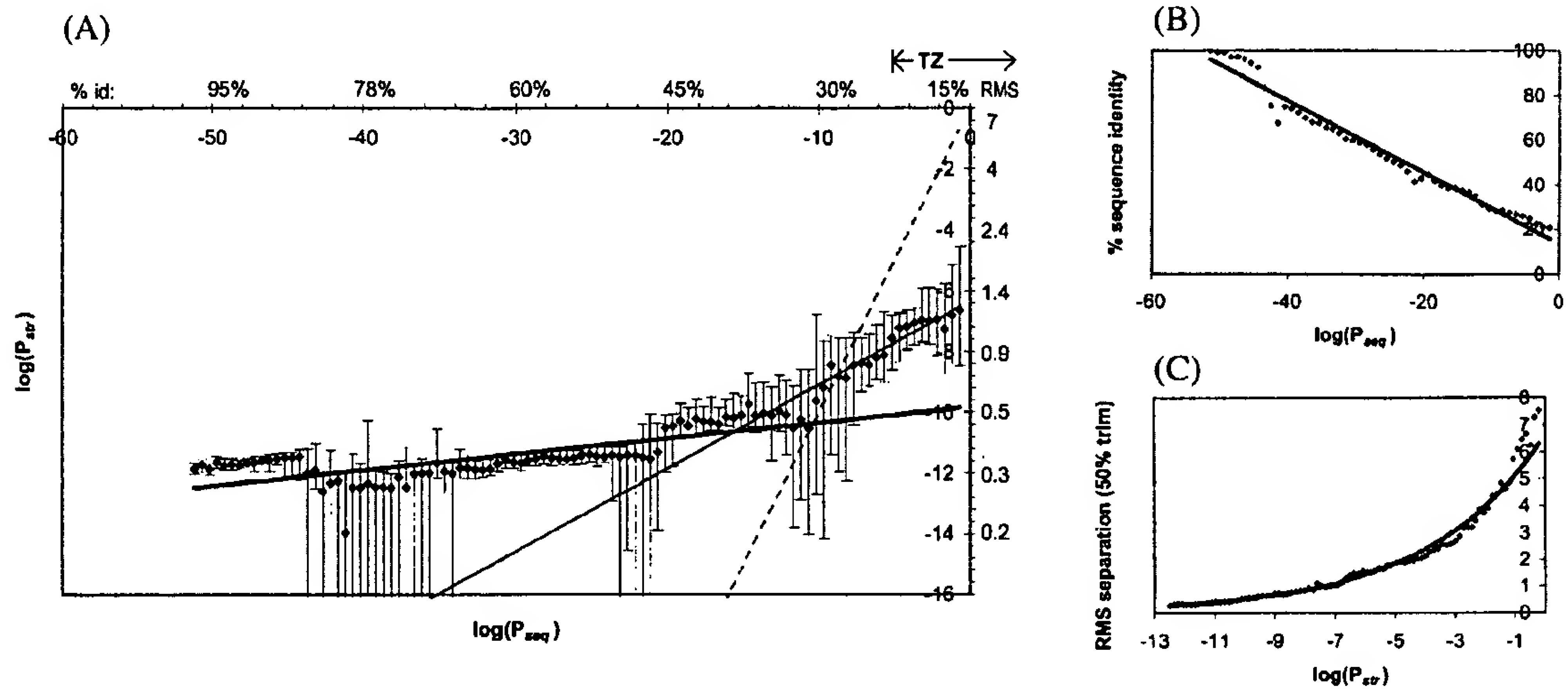


Figure 4. Probabilistic scores: *P*-values. P_{seq} and P_{str} are *P*-values calculated from S_{seq} and S_{str} according to the formalism given by Levitt & Gerstein (1998). Both quantities have the same overall functional form in terms of an extreme value distribution:

$$P = 1 - \exp(-\exp(-Z))$$

where P is either P_{seq} or P_{str}. For P_{seq}, $Z = S_{seq}/a - 2\ \ln M - b/a$, where $a = 5.84$, $b = -26.3$, and M is the geometric mean of the lengths of the two sequences (i.e. $M^2 = nm$, where n and m are the two sequence lengths). For P_{str}, Z is a function of S_{str} and N, the number of matched residues: For $N < 120$:

$$Z = (S_{str} - c\ln^2 N - d\ln N - e)/(f\ln N + g)$$

For $N \geqslant 120$:

$$Z = (S_{str} - a\ln N - b)/(f\ln 120 + g)$$

At $N = 120$, continuity implies that:

$$a\ln 120 + b = c\ln^2 120 + d\ln 120 + e \quad \text{and} \quad a = 2c\ln 120 + d$$

This, in turn, allows the calculation of the constants:

$$a = 171.8,\ b = -419.4,\ c = 18.4,\ d = -4.50,\ e = 2.64,\ f = 21.4,\ g = -37.5$$

(a) of this Figure is analogous to Figures 3(a) and 2(b), with the exception of the fits. It is a log-log (base 10) plot relating P_{seq} and P_{str}. We show the median log(P_{str}) value for each log(P_{seq}) bin, along with quartile bars above and below. We have added approximate percent identity and RMS values to the x and y axes to aid interpretation of the graph in terms of more familiar scores. The values were calculated using the calibration curves in (b) and (c). The straight-line nature of the log-log plot reveals distinct relations inside and outside the twilight zone, labeled TZ. (The area of percent identity below the twilight zone does not appear in P_{seq} graphs, there is no significance for such low sequence similarity; thus all data points in that zone appear at $P_{seq} = 1$ or $\log[P_{seq}] = 0$.) The thick line in the figure is fit to the median P_{str} values for P_{seq} values outside the twilight zone; its equation is:

$$P_{str} = 10^{-10} P_{seq}^{0.05}$$

The thin line is fit to the data inside the twilight zone; it follows the relation:

$$P_{str} = 10^{-6} P_{seq}^{0.274}$$

For reference we include the dotted line, representing the function $P_{str} = P_{seq}$, where sequence and structural similarity are equally significant. See the text for a discussion of how the two trends might be interpreted with respect to this line.

can be used to transfer precisely structural annotation based on the degree of sequence homology. In genome analysis, however, one is usually more interested in finding a functional annotation for an open reading frame based on similarity to well-known proteins; yet the sequence-function and structure-function relationships have not been as explicitly characterized. The fundamental obstacle to extending this and similar investigations to deal with function is the absence of a clear measure of

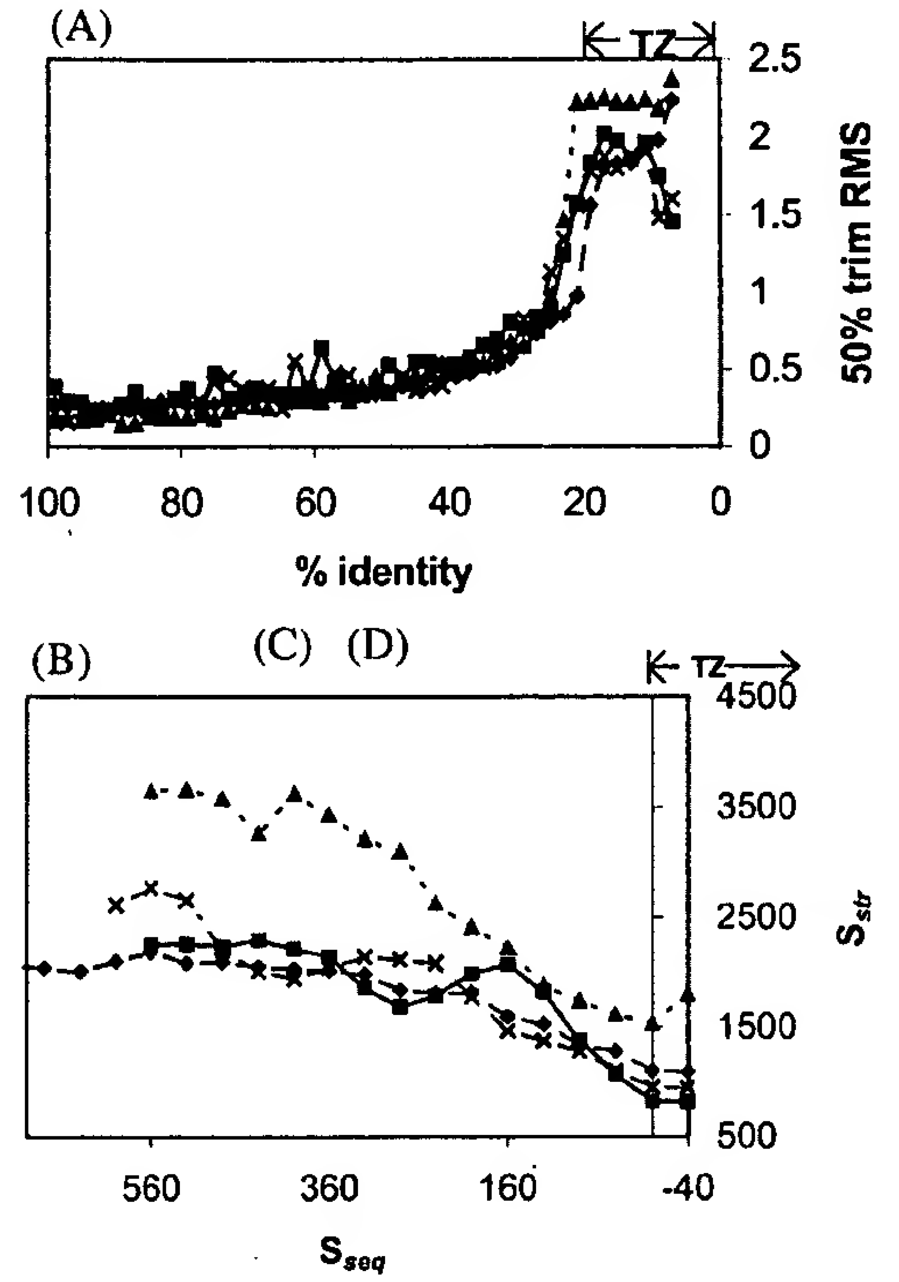


Figure 5. SCOP class differences. Previously it has been observed that secondary structural composition does not cause deviations from the trends in structure and sequence similarity (Flores *et al.* 1993). To test this observation we looked at the scores divided by SCOP class. The following legend applies to the graphs: (—■—), all alpha; (- ◆ -), all beta; (-- ▲ --), alpha/beta; (-- × --), alpha + beta. (a) Median RMS values for each percent identity bin. The traditional scores reveal no dependency on class. However, in (b) α/β pairs consistently score higher S_{str} scores than pairs in other classes. This is a consequence of the dependence of S_{str} on length; domains in the α/β class are longer, on average, than in the other classes.

functional similarity. Although we were able to present three different quantitative measures of structural relatedness, an analogous situation for function does not exist. How can one express quantitatively the degree of similarity between a triosephosphate isomerase and a glucose-6-phosphate isomerase? How do they compare to trp repressor?

The absence of a clear measure of functional similarity is not the only obstacle in transferring the functional annotations between proteins with different degrees of homology. The definition of function itself is often vague. More specifically, at present there is an absence of such important information as a standardized vocabulary for protein functional annotations with an associated numbering scheme, descriptions of monomer functions of subunits of multisubunit proteins and hierarchical functional assignments for proteins with multiple functions. As a consequence of these difficulties there is no functional equivalent to the hierarchical fold classification for domains in PDB.

As signs of progress in this direction, several functional classifications have been developed to date. One is the ENZYME system developed by the Enzyme Commission (EC) to classify enzymes by reaction type (Webb, 1992). This system has the advantage that it is "universal," applicable to proteins in many different organisms, and is in wide use. However, it also has several drawbacks. First of all, it does not consider catalytic reaction mechanisms (Riley, 1998a), often ignoring obvious similarities. Second, it presumes a 1:1:1 relationship between gene, protein and reaction, although this is often not the case (an enzyme can have two functions, or two polypeptides from two different genes can oligomerize to perform a single function). Perhaps the most significant drawback of the EC classification is that it applies to only enzymes.

A number of more comprehensive schemes have been developed, which classify non-enzymes as well as enzymes. Most of these focus on individual organisms. Several such schemes exist, for instance, GenProtEC/EcoCyc for *E. coli* (Karp *et al.*, 1998b; Riley & Labedan, 1996; Riley, 1998b), MIPS for yeast (Mewes *et al.*, 1998), Ashburner's functional classification for Drosophila, which is connected to FLYBASE (Ashburner & Drysdale, 1994), and EGAD for human ESTs (Adams *et al*, 1995). These classifications possess some advantages. They have additional levels of hierarchy that help present a more comprehensive picture of genotype-phenotype relationships. On the other hand, these classifications still leave much room for improvement. For example, there is no standardized vocabulary to allow for keyword searches among multiple databases and across organisms, and there are inconsistencies in category numbering style.

Finally, there has been some promising work going beyond the ENZYME and organism-focused classifications. There has been progress on completely automated functional classification (des Jardins *et al.*, 1997; Tamames *et al.*, 1997), which has the potential for putting function assignments on a more objective basis. There are a number of databases synthesizing the various enzyme functions into coherent pathways and systems (e.g. KEGG and WIT, Ogata *et al.*, 1999; Selkov *et al.*, 1998). There also have been some very recent attempts to develop cross-species classifications of non-enzyme functions in the framework of the Gene Ontology Project (GO, geneontology.org). GO is a joint project between FlyBase, the Saccharomyces Genome Database and Mouse Genome Informatics, attempting to merge the fly, yeast and mouse functional classification schemes. However, a truly universal system for classifying all protein functions in all organisms within the same framework remains quite a challenge because of the

sheer diversity of organisms and distinct protein functions.

Our simple functional classification of SCOP domains: FLY+ENZYME

Given the discussed limitations, we constructed a simple functional classification for the SCOP domains included in our comparison; our classification is based on a merger of two of the existing functional annotations and a cross-referencing of subsets of this combination with some of the organism-specific schemes. First, we used pairwise comparison to cross-reference the PDB domains against the Swissprot database (Bairoch & Apweiler, 1998), as described by Hegyi & Gerstein (1999). We chose to assign protein functions according to Swissprot because it provides more comprehensive functional annotations than SCOP.

We were initially able to divide all entries into enzymes and non-enzymes, a division that represents the highest level of functional difference in our classification scheme (Figure 6). For the enzyme category, we transferred EC (Webb, 1992) numbers to those SCOP domains with a one-to-one match to a Swissprot enzyme. Only one-to-one matching entries could be considered because Swissprot assigns ENZYME numbers to entire proteins, whereas SCOP is a domain-based classification; therefore we could be confident about the classification of only those domains which map to an entire Swissprot entry.

In the absence of an EC-type classification for non-enzymes, we assigned functions to non-enzymatic SCOP domains according to Ashburner's original classification of Drosophila protein functions. This classification is derived from a controlled vocabulary of fly terms. It is available on the web and loosely connected with the FLYBASE database (Ashburner & Drysdale, 1994). For clarity, we precisely describe the specific files and version (1.55, 1997) of the classification that we used in the caption to Figure 6, and we will hereafter refer to these data files as constituting the original FLY classification.

The FLY classification is a dynamic object, changing as more is learned about the fly and other organisms. This is particularly true of late with the imminent completion of the Drosophila genome. In fact, since the completion of our analysis, the FLY classification has been superceded by the new GO classification (see above).

The hierarchical structure of the FLY classification makes it well suited for classifying non-enzymatic SCOP entries in a manner comparable to the ENZYME assignments for the enzymes. Another advantage of this classification is that it is more compatible with the makeup of the PDB than the *E. coli* and yeast classifications, as Drosophila is a multi-cellular organism, and many of the known structures come from animals. We were able to use the original FLY classification as a framework to which we added functional categories and individual proteins. For instance, we added "Hemoglobin" to the "Physiological Processes - Respiration" category. Another example is the "Physiological processes - Immunity" category (Figure 6(b)), to which we added immune system proteins. Many of the additions would not be necessary in the context of the new cross-species GO system. We also modified slightly the numbering scheme in the original FLY classification in order to assign a unique hierarchical number to each protein domain (Figure 6(b)). We will refer to our augmented FLY classification as the FLY+ scheme, and our merged scheme as the FLY+ ENZYME classification.

As discussed earlier, the universal functional classification of proteins is very challenging and may not be possible with the current level of knowledge about genes, proteins and genomes. Consequently, the FLY + ENZYME classification of SCOP proteins is somewhat incomplete and inconsistent and retains many of the limitations of its components (Hegyi & Gerstein, 1999; Riley, 1998a). It is not yet broad enough to include many plant, virus and bacterial proteins. Nevertheless, it was sufficient for our analysis, as we were able to classify a very large number of the total 30,000 pairs.

Determining functional similarity

Using our compound functional classification, we were able to assign a level of functional similarity to each domain pair. According to our scheme, a pair can have no functional similarity (an enzyme paired with a non-enzyme) or it can have one of three levels of similarity:

(i) General similarity. Both domains are enzymes or both are non-enzymes.

(ii) Same functional class. Both domains share the first component of their ENZYME or FLY + numbers, e.g. 1.1.1.1 alcohol dehydrogenase and 1.3.1.1 cortisone beta-reductase (for enzymes), or 3.3.2.1.2 calcicyclin and 3.6.3.2.1 calmodulin (for non-enzymes).

(iii) Same precise function. Both domains share three components of their ENZYME or FLY + number, e.g. 1.1.1.1 alcohol dehydrogenase and 1.1.1.3 homoserine dehydrogenase (for enzymes) or 1.2.9.1.1.1 Arc repressor and 1.2.9.1.1.1 C-jun (for non-enzymes; both are transcription factors). A pair that shares precise function must also, by definition, share functional class and general similarity.

Based on those assignments we calculated the percentage of total pairs at a given level of sequence or structural similarity possessing each level of functional similarity. The results appear in Figure 7.

Sequence and function

The relation between sequence similarity and functional similarity behaves as one might expect, with sigmoidal curves that drop off sharply at particular conservation thresholds, and with the three levels of functional similarity (precise function, functional class and general similarity) having progressively lower thresholds. Figure 7(a) shows that precise function is not conserved below 30-40% sequence identity, whereas functional class is conserved for sequence identities as low as 20-25%. Below 20%, general similarity is no longer conserved; among pairs of approximately 7% sequence identity, about 40% are enzymes paired with non-enzymes. It is important to note that in all the pairs considered here, the domains share the same fold. Functional similarity at low percent identities (e.g. 7%) would be much less for all possible pairs of domains rather than just for those with the same fold. It is also important to remember that our thresholds for functional conservation are statistical averages over many sequences; one will, of course, be able to find individual cases that diverge more or less rapidly.

There are differences between the functional conservation thresholds of enzymes and non-enzymes, with enzymes appearing to more highly conserve precise function than non-enzymes, but non-enzymes conserving functional class more highly than enzymes. This may reflect that in our classification, the non-enzyme functional classes are broader and hence easier to conserve than those of the enzymes, while the non-enzymatic precise functions are more specific.

When P_{seq} is used as the measure of sequence similarity (Figure 7(b)) the results look somewhat different, it appears that functional class is conserved for the entire range of sequence similarities. In this case, percent identity is actually more discriminating than P_{seq} because functional class diverges only at sequence similarities that are low enough that they have little or no statistical significance, i.e. for P_{seq} the divergence is compressed near the vertical axis of the graph.

Structure and function

The relation between similarity in structure and function is somewhat less straightforward than that between similarity in sequence and function. Figure 7(c) shows the relationship between RMS and functional similarity. Broadly, it appears similar to that for percent identity and functional similarity; however, the thresholds for conservation of the various types of functional similarity are less sharp.

RMS is more revealing with respect to functional similarity than the non-traditional structural scores, S_{str} and P_{str}. (Data for S_{str} and P_{str} are not shown but are available from the website.) The reason is that, while very structurally similar pairs all have RMS scores clustered between 0 and 0.5 Å, S_{str} has a large range of scores for similar pairs due to the length dependency, and P_{str} does not have any limit for maximum similarity. The wide range of possible S_{str} and P_{str} scores for similar structures tends to blur the broad sigmoid curves so much so that they are no longer apparent.

Alternative functional classifications: MIPS and GenProtEC

To get some perspective on the degree to which our results reflected the particularities of our combined FLY + ENZYME classification, we decided to try the same comparisons based on the well-known functional classifications for yeast and *E. coli*, MIPS and GenProtEC (Mewes *et al.*, 1998; Riley & Labedan, 1996; Riley, 1998b). These classifications have the advantage that they integrate enzyme and non-enzyme functions from the start and are widely used. However, as they are only applicable to individual organisms, we could only use them to classify a considerably smaller subset of the known structures than the compound FLY + ENZYME system.

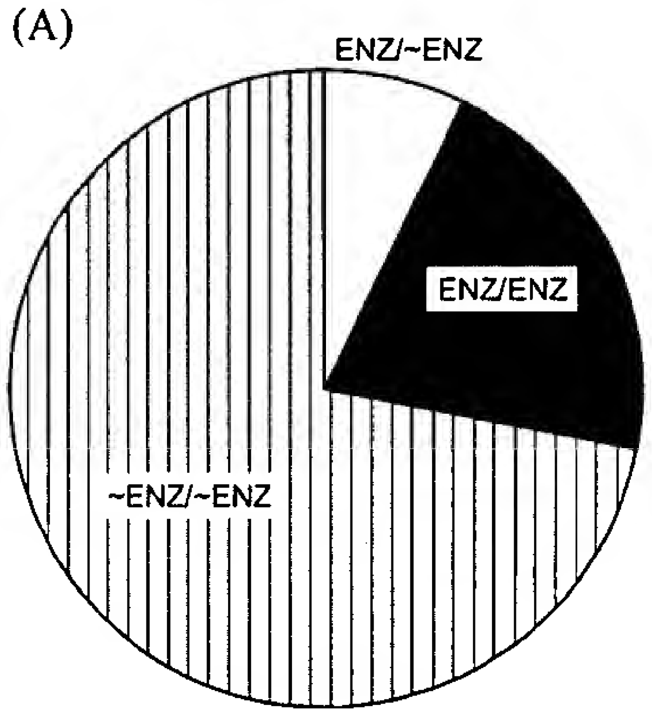
The specific way we used the MIPS and GenProtEC classifications to assign function to structures and to calculate functional similarities is described in the legend to Figure 7. Our results in terms of functional conservation (precise and class) at various levels of percent identity are shown in Figure 7(d). We observe the same general relationships as we did for our FLY-+ENZYME scheme. That is, the functional conservation curves have a sigmoidal shape and have cut-offs for precise functional similarity after 40% and for functional class similarity at lower values. However, because the MIPS and GenProtEC classifications are restricted to individual organisms, each curve represents considerably fewer data points than do the curves based on the FLY + ENZYME scheme; this required us to "bin" the MIPS and GenProtEC curves in a somewhat coarser fashion.

Discussion and Conclusion

Here, we assessed the transfer of functional and structural annotation by analyzing the relationships between similarity in sequence, structure and function. The ~30,000 protein domain pairs of varying levels of similarity (at least the same fold) that we constructed out of the SCOP classification show quantitative sequence-structure relationships consistent with previous research. The exponential relationship is consistent across the secondary-structural classes and holds for newer probabilistic scoring methods.

The sequence-function and structure-function relationships have not been studied as precisely due to the lack of a robust functional classification and measure of functional similarity. To overcome

(A)



(B)

Category type	Specific example	Original	Modified
Function primitive	Cell structure	0.	3
Process or Molecule	Nuclcus	1.	3.1
Pathway or Macromolecule	Nuclear Membrane structure	2.	3.1.1
Individual instances	Nuclear envelope protein	3.	3.1.1.1
Individual instances	Lamin	4.	3.1.1.1.1
Individual instances	LaminA	5.	3.1.1.1.1.1
	...		
Process or Molecule	Nucleocyt. transport	2.	3.1.2
Individual instances	Nuclear pore structure	3.	3.1.2.1
	...		
Process or Molecule	Cell motility	1.	3.12
Pathway or Macromolecule	Motor protein	2.	3.12.1
Individual instances	Myosin	3.	3.12.1.1
Process or Molecule (new)	Immunity	1.	5.6
Pathway or Macromolecule (new)	Immunoglobulin	2.	5.6.1
Individual Instances(new)	Immunoglobulin binding prot.	3.	5.6.1.1

(C)

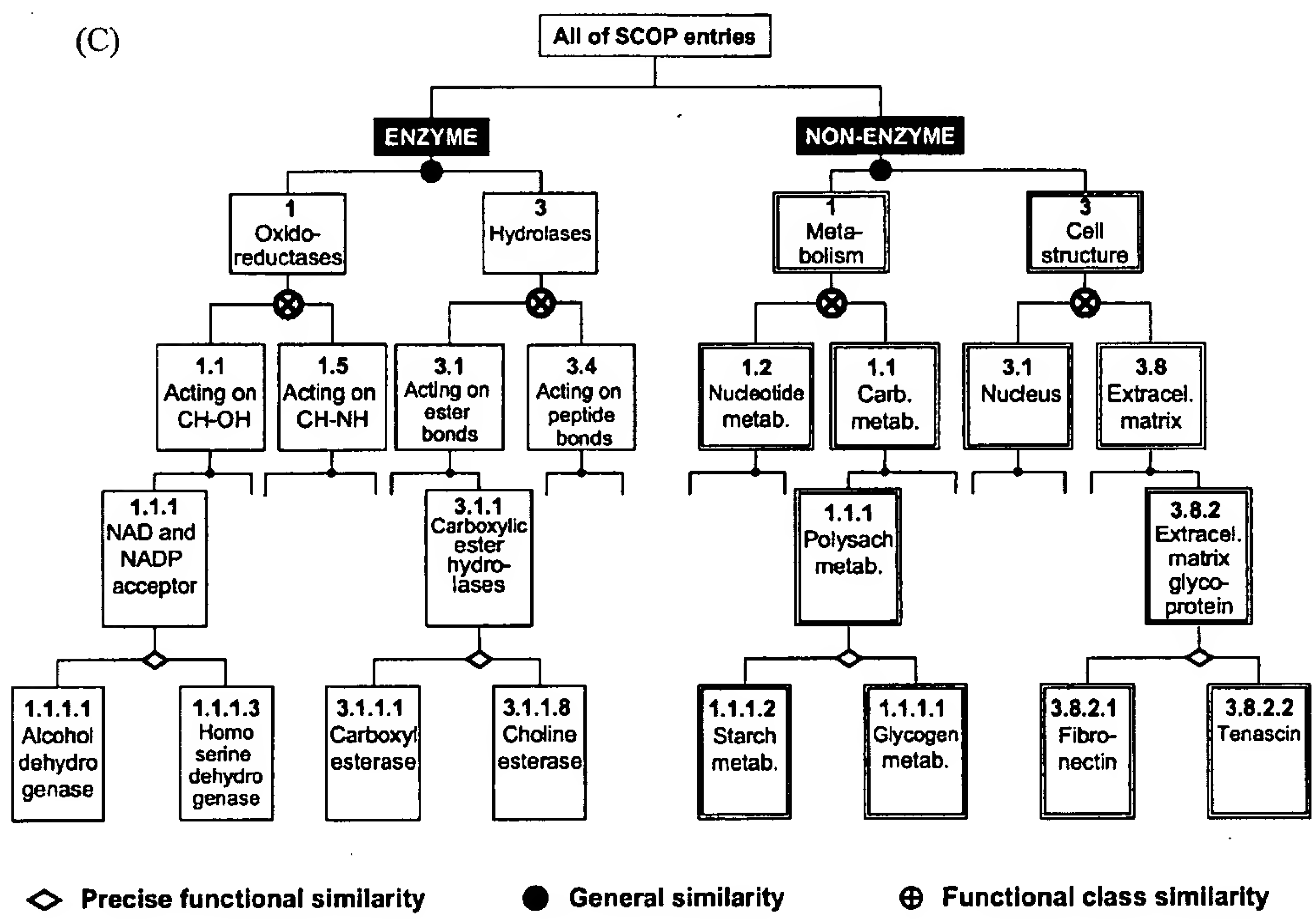


Figure 6. Functional classification of enzymes and non-enzymes. (a) Divides the pairs by general function. There are three categories of pairs: (i) enzymes paired with non-enzymes (no general functional similarity), labeled ENZ/~ENZ; (ii) enzymes paired with enzymes (same general function), labeled ENZ/ENZ; and (iii) non-enzymes paired with non-enzymes (same general function). Pairs for which one or both domains could not be identified as enzyme or non-enzyme are not included in this chart. Enzymes are classified according to the EC system (Webb, 1992). The first component of the number represents the nature of reaction and is called class. There are six classes: oxidoreductases, transferases, hydrolases, lyases, isomerases and ligases. The next level is subclass. It refers to the chemical groups on which the enzyme acts. For example, the first class, oxidoreductases, has 19 subclasses that are arranged according to the donor group that undergoes oxidation (CH-OH, aldehyde or oxo group, CH-CH group, etc). For another group of enzymes (hydrolases) subclass is determined by the nature of the bond: ester bond, peptide bond, etc. The next level is sub-subclass. For oxidoreductases this indicates the acceptor group: NAD(+) and NADP(+), or cytochrome; for hydrolases the sub-subclass represents the nature of substrate (carboxylic ester hydrolases, thiolester hydrolases, etc.). The fourth level represents a unique number for each individual enzyme, for example, 1.1.1.1: alcohol dehydrogenase. (b) Shows how we adapted the functional classification of Drosophila gene products developed by M. Ashburner. This classification is loosely connected with FLYBASE (Ashburner & Drysdale, 1994). We used version 1.55 (4 August 1997) that was available from Ashburner's website:

http : //www.ebi.ac.uk/ ~ ashburn

The specific files that we used were taken from the ftp directory:

ftp.ebi.ac.uk/databases/edgp/misc/ashburner

this we constructed our own classification by merging and extending the ENZYME and FLY schemes and assigning levels of functional similarity. Our measures of functional similarity provide curves relating function to sequence and structure; when relating functional conservation to sequence divergence, we find distinct thresholds at ~40% for precise function and ~25% for functional class.

One of the interesting results that emerges from this is that percent identity is more useful for quantifying functional divergence than the newer probabilistic scores. In general, modern probabilistic scores, such as P_{seq}, are better at discriminating amongst highly diverged sequences (near the twilight zone) than percent identity, since they better take into account gaps and conservative substitutions (of similar amino acids). However, for very similar pairs of sequences, percent identity is a simpler and more direct measure of divergence (essentially a Hamming distance). Since divergence in precise function takes place before that in structure (well before the twilight zone), it is quite reasonable that percent identity is more successful at measuring the former than the latter and that the converse is true for the probabilistic scores. In other words, percent identity is better calibrated for discriminating amongst very close, significant relationships and P_{seq} for more distant ones.

Practical implications

The sequence-structure and sequence-function relationships described here provide practical information for genome annotation in terms of folds and functions. Table 1 summarizes the relative advantages of the different scoring methods we used. Using the trends in sequence and structure similarity, one can assess the degree to which structural annotation can be transferred between sequences at a given level of sequence similarity. The sequence and function similarity thresholds potentially establish minimum requirements of sequence similarity for reliable function prediction. Note that because the protein domain pairs considered here all share the same fold, the numbers for all possible pairs will differ in the region of very little sequence identity, in which the sequence similarity is not enough to indicate the same fold.

We refer to these as constituting the original FLY classification. Recently, the FLY classification has been superceded by the GO (Gene Ontology) Project classification, which merges fly, mouse and yeast annotation. Files related to the GO classification are available from www.geneontology.org In the original FLY classification all members of the highest level are labeled 0, representatives of the next level are labeled 1, and all lower levels are labeled 2 through to 9. We changed the numbering scheme so that it will reflect the hierarchical nature of the classification. This Figure illustrates sections of the original and modified classification. The top level in the FLY classification scheme is called "Function primitive" (level 0) and includes five classes: "Metabolism," "Intracellular protein traffic," "Cell structure," "Developmental process," "Physiological process," and "Behavior." The next level after "Function primitive" is "Process" or "Molecule" (level 1 in Ashburner's classification). For "Function primitive - Metabolism" the processes are "Carbohydrate metabolism," "Nucleotides and nucleic acids metabolism," etc. For "Function primitive - Cell Structure" the "Process" can be "Nucleus," "Mitochondrion," "Membrane," etc. The next level is "Pathway" or "Macromolecule" (level 2 in the original classification). "Pathway" can include "Metabolic pathway," "Signaling pathway," or "Developmental pathway." The "Macromolecule" category includes "Protein" and "Nucleic Acid". We added categories to the original classification in order to classify some mammalian proteins that are widely represented in SCOP but are absent from the original FLY scheme. These categories include immune system proteins (labeled "new" in (b) and respiratory proteins such as hemoglobin and myoglobin that we added to "Function primitive - Physiological process - Respiration". We call our adaptation of the original FLY scheme, FLY + . Further information on this adaptation is available at:

http : //bioinfo.mbb.yale.edu/align/func

(c) The overall hierarchy of our final scheme and identification of the different levels of similarity. If two proteins are both enzymes or both non-enzymes, then they possess general functional similarity. If they share the first component of their classification numbers, then they are in the same functional class. If they share the first three components of their enzyme numbers (or the equivalent for non-enzyme numbers, depending on category) then they have the same precise function. A significant difference between the two main branches of the hierarchy is that the levels of the ENZYME classification do not correspond exactly to those in the FLY+ system because the fly classification is more extensive than the enzyme classification. For instance, the FLY classification takes into account aspects of cellular (cytoskeleton, metabolic pathways, etc.) and phenotypic function (morphology, physiology, behavior) that are absent from the ENZYME scheme. This makes our classification of SCOP proteins somewhat unbalanced, as non-enzymes have much broader and more loosely defined functional classes. As a consequence, while each enzyme is assigned a four-component number, the length of a non-enzyme number varies, depending on the functional category to which it belongs. For example, myosin is assigned a number that happens to have the same length as EC numbers: 3.12.1.1. However, transcription factors are numbered 1.12.9.1.1.1. We took into account this varying hierarchy depth in deciding how many components are necessary to identify precise function in each category. Note that what we mean by domains having the same precise function is not the same as the domains coming from the same essential protein.

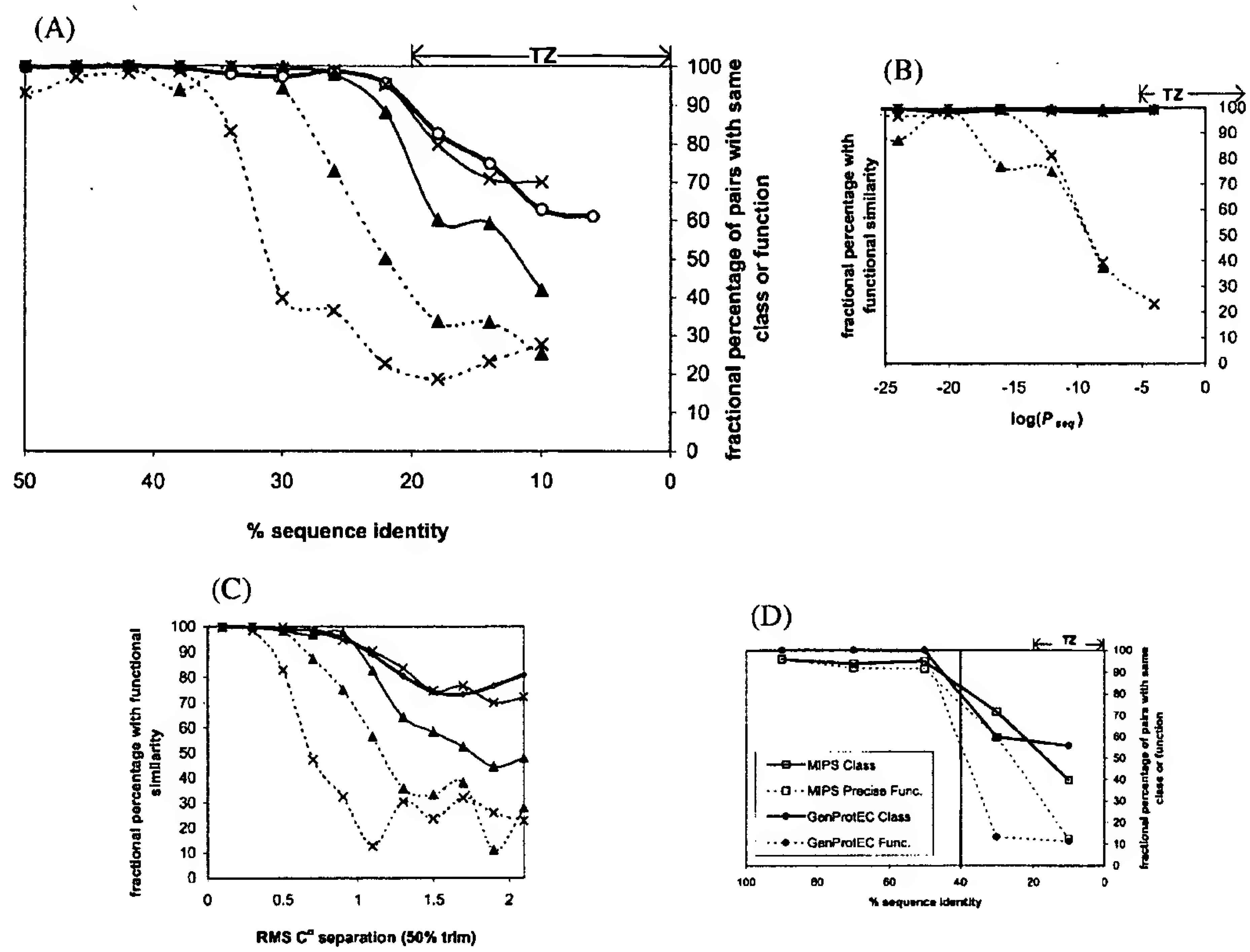


Figure 7. Linking sequence, structure and function. We express functional similarity as the fractional percentage of pairs at a given level of sequence/structural similarity for which the paired domains share a precise function, functional class, or general similarity (according to our classification, see Figure 6). The following legend applies to (a) through (c): (—O—), general similarity; (— x —), non-enzymes with same functional class; (—▲—), enzymes with same functional calss; (--- x ---), non-enzymes with same precise function; and (---▲---), enzymes with the same precise function. (a) Relates functional similarity to sequence similarity in terms of percent identity. The functional similarity appears as a sharp sigmoid, with distinct thresholds of divergence for precise function, functional class, and general similarity. Enzymes are paired with non-enzymes only at very low percent identity, in and below the twilight zone (labeled TZ). At slightly higher sequence identity, pairs diverge with respect to functional class, and beyond 40% identity with respect to precise function. Note that 50-100% identity is not shown because almost all domains that are that similar share function with their counterparts. (b) Shows the same data using P_{seq} as the measure of sequence similarity. Only the divergence in precise function is visible because there is such little significance for the low sequence similarity at which functional class and general similarity diverge, all data points in that region appear near $P_{seq} = 1$ or $\log[P_{seq}] = 0$ (the y-axis). (c) Illustrates that the structure-function relation is not as clearly defined as that for sequence and function. Functional similarity expressed in terms of RMS separation appears as a broad sigmoid curve; there are thresholds of divergence for precise function, but the divergences in functional class and general similarity are more gradual. The thresholds are apparent only because RMS clusters the most structurally similar pairs between scores of 0 and 0.5 Å. For this reason, RMS is better at discerning functional similarity than S_{str} and P_{str}, which do not cluster the most similar pairs around a set limit. (d) Shows the same relationships (functional conservation *versus* percent identity) as in (a), except that for this graph functional similarity is determined in terms of the MIPS (Mewes *et al.*, 1998) and GenProtEC (Riley, 1998b) classifications rather than the FLY + ENZYME scheme. The legend appears as the inset on the graph. We assigned MIPS and GenProtEC classifications to SCOP domains based on sequence comparisons to classified yeast and *E. coli* open reading frames (ORFs), respectively. The SCOP domain most closely matching each ORF classified in MIPS or GenProtEC was assigned the corresponding MIPS or GenProtEC function number. Only matches of 80% sequence identity or greater were considered. We used this SCOP domain as a functional representative; when determining functional similarity, we assigned to SCOP domains with no MIPS or GenProtEC functional designation the function of the closest representative with at least 85% sequence identity, if one existed. GenProtEC functional identifiers are three-component numbers. We consider a pair of domains sharing the first component of their functional designation to be in the same functional class. Domains that share all three components are said to have the same precise function. For MIPS the functional designation is not as straightforward, as one ORF can be assigned multiple functions. Therefore we consider domains which have at least one function in common to share functional class. Domains with all functions in common, the same combination of identifiers, share precise function. Because MIPS and GenProtEC each classify the proteins of a single organism, yeast and *E. coli*, respectively, these classifications can determine the functional similarities of only a small fraction of all our SCOP domain pairs. The data based on these classifications, appearing in (d), are therefore very sparse compared to the data in (a)-(c). Despite the coarseness of the data, functional similarity based on the MIPS and GenProtEC classifications follows the same general relation to sequence similarity as does functional similarity based on the more comprehensive FLY + ENZYME scheme. Vertical line indicates an approximate threshold of functional divergence at 40% identity.

Table 1. Summary of scoring methods

	Sequence similarity	Structural similarity	Features	Limitations
Traditional scores	Per cent sequence identity	RMS C^{α} separation	Well understood, in use; percent identity better for looking at functional similarity	RMS depends most highly on worst matches, requiring arbitrary trimming; percent identity is insensitive to gaps and conservative substitutions
Alignment similarity scores	S_{seq}	S_{str}	Analogous similarity scores, S_{str} depends most highly on best matches	Dependence on alignment length
Modern probabilistic scores	P_{seq}	P_{str}	Statistical significance, unified framework for different comparisons	Not as familiar as RMS and percent identity

The Table lists the schemes presented here for characterizing the sequence-structure relationship, along with their relative advantages and disadvantages.

Practically, then, when one searches an uncharacterized open reading frame against known structures, if the open reading frame matches a structure with a good *e*-value or percent identity, then the curves presented here can be used to check how the functional and detailed structure annotation will transfer. For example, if an unknown open reading frame matches a PDB structure with an *e*-value of 0.001 and a percent identity of 30%, then one can be assured that it has the same fold (Brenner *et al.*, 1998) and according to our analysis it has a two-thirds chance of having the same exact function. Furthermore, it has a ~99% chance of having the same functional class and its structure probably diverges from the known structure by a trimmed RMS of less than 0.7 Å.

Future directions

There are a number of directions in which we might extend this analysis. With respect to the sequence-structure relation, we can reduce the overrepresentation of the immunoglobulins and improve the calculation of P_{str} (by redoing the fit to the extreme value distribution reported by Levitt & Gerstein (1998) to eliminate residual length-dependency.

In the functional realm, we can investigate if and how the sequence-function and structure-function relationships vary for different categories of proteins. For example, although we found consistency of the sequence-structure relationship among secondary structural classes, Hegyi & Gerstein (1999) found that the distribution of enzymes and non-enzymes varies with secondary structural class. A related issue is that of conformational changes. It is conceivable that among domains with very similar sequences but structures that differ by a conformational change, function is less conserved than it is among similar sequences with more similar structures.

Perhaps the most important direction in which to further this work is the augmentation of the functional classification. With the growing amount of fully sequenced genomes there is a need for the development of a comprehensive system for functionally classifying proteins, a complete classification for the entire universe of protein functions. It will be a difficult process, as many existing organism-specific classifications will have to be merged, but the end result will have the advantage of not being biased towards any one organism. Such a universal classification will allow much more reliable transfer of functional annotation.

Acknowledgments

We thank A. Lesk for helpful conversations and supplying us with reference data for Figure 2, S. Brenner for providing carefully curated SCOP domain sequences, and H. Hegyi, W. Krebs and V. Alexandrov for assistance with the sequence comparisons, development of the FLY + ENZYME scheme, and design of the web database. M.G. thanks the Keck and Donaghue foundations for financial support.

References

Abagyan, R. A. & Batalov, S. (1997). Do aligned sequences share the same fold? *J. Mol. Biol.* **273**, 355-368.

Adams, M. D., Kerlavage, A. R., Fleischmann, R. D., Fuldner, R. A., Bult, C. J., Lee, N. H., Kirkness, E. F., Weinstock, K. G., Gocayne, J. D., White, O., Venter, J. C., *et al.* (1995). Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature*, **377**, 3-174.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tools. *J. Mol. Biol.* **215**, 403-410.

Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994). Issues in searching molecular sequence databases. *Nature Genet.* **6**, 119-129.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation

of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.

Andrade, M. A. & Sander, C. (1997). Bioinformatics: from genome data to biological knowledge. *Curr. Opin. Biotech.* **8**, 675-683.

Ashburner, M. & Drysdale, R. (1994). Flybase: the *Drosophila* genetic database. *Development*, **120**, 2077-2079.

Attwood, T. K., Flower, D. R., Lewis, A. P., Mabey, J. E., Morgan, S. R., Scordis, P., Selley, J. N. & Wright, W. (1999). PRINTS prepares for the new millennium. *Nucl. Acids Res.* **27**, 220-225.

Bairoch, A. & Apweiler, R. (1998). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucl. Acids Res.* **26**, 38-42.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.

Bork, P. & Koonin, E. V. (1996). Protein sequence motifs. *Curr. Opin. Struct. Biol.* **6**, 366-376.

Bork, P., Ouzounis, C. & Sander, C. (1994). From genome sequences to protein function. *Curr. Opin. Struct. Biol.* **4**, 393-403.

Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M. & Yuan, Y. (1998). Predicting function: from genes to genomes and back. *J. Mol. Biol.* **283**, 707-725.

Brenner, S. E. (1999). Errors in genome annotation. *Trends Genet.* **15**, 132-133.

Brenner, S. E., Chothia, C., Hubbard, T. J. & Murzin, A. G. (1996). Understanding protein structure: using SCOP for fold interpretation. *Methods Enzymol.* **266**, 635-643.

Brenner, S. E., Chothia, C. & Hubbard, T. J. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA*, **95**, 6073-6078.

Bryant, S. H. & Altschul, S. F. (1995). Statistics of sequence-structure threading. *Curr. Opin. Struct. Biol.* **5**, 236-244.

Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823-826.

Chothia, C. & Lesk, A. M. (1987). The evolution of protein structures. *Cold Spring Harbor Symp. Quant. Biol.* **52**, 399-405.

des Jardins, M., Karp, P. D., Krummenacker, M., Lee, T. J. & Ouzounis, C. A. (1997). Prediction of enzyme classification from protein sequence without the use of sequence similarity. *ISMB*, **5**, 92-99.

Doolittle, R. F. (1987). *Of Urfs and Orfs*, University Science Books, Mill Valley, CA, USA.

Enright, A. J., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86-90.

Fetrow, J. S. & Skolnick, J. (1998). Method for prediction of protein function from sequence using the sequence to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T_1 ribonucleases. *J. Mol. Biol.* **281**, 949-968.

Fetrow, J. S., Godzik, A. & Skolnick, J. (1998). Functional analysis of the *Escherichia coli* genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *J. Mol. Biol.* **282**, 703-711.

Flores, T. P., Orengo, C. A., Moss, D. S. & Thornton, J. M. (1993). Comparison of conformational characteristics in structurally similar domain pairs. *Protein Sci.* **2**, 1811-1826.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Venter, J. C., *et al.* (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science*, **270**, 397-403.

Fraser, C. M., Norris, S. J., Weinstock, G. M., White, O., Sutton, G. G., Dodson, R., Gwinn, M., Hickey, E. K., Clayton, R., Ketchum, K. A., Sodergren, E., Hardham, J. M., McLeod, M. P., Salzberg, S., *et al.* (1998). Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science*, **281**, 375-388.

Gerstein, M. (1997). A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J. Mol. Biol.* **274**, 562-576.

Gerstein, M. (1998a). Measurement of the effectiveness of transitive sequence comparison, through a third 'intermediate' sequence. *Bioinformatics*, **14**, 707-714.

Gerstein, M. (1998b). Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins: Struct. Funct. Genet.* **33**, 518-534.

Gerstein, M. (1998c). How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Folding Des.* **3**, 497-512.

Gerstein, M. & Altman, R. (1995). Average core structures and variability measures for protein families: application to the immunoglobulins. *J. Mol. Biol.* **251**, 161-175.

Gerstein, M. & Hegyi, H. (1998). Comparing microbial genomes in terms of protein structure: surveys of a finite parts list. *FEMS Microbiol. Rev.* **22**, 277-304.

Gerstein, M. & Levitt, M. (1996). Using iterative dynamic programming to obtain accurate pairwise and multiple alignments of protein structures. *ISMB*, **4**, 59-67.

Gerstein, M. & Levitt, M. (1998). Comprehensive assessment of automatic structural alignment against a manual standard, the SCOP classification of proteins. *Protein Sci.* **7**, 445-456.

Hegyi, H. & Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288**, 147-164.

Heinikoff, S. & Heinikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915-10919.

Hubbard, T. J. P., Murzin, A. G., Brenner, S. E. & Chothia, C. (1997). SCOP: a structural classification of proteins database. *Nucl. Acids Res.* **25**, 236-239.

Karlin, S. & Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA*, **87**, 2264-2268.

Karlin, S. & Altschul, S. F. (1993). Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl Acad. Sci. USA*, **90**, 5873-5877.

Karlin, S., Bucher, P., Brendel, V. & Altschul, S. F. (1991). Statistical methods and insights for protein and DNA sequences. *Annu. Rev. Biophys. Biophys. Chem.* **20**, 175-203.

Karp, P. D. (1996). A protocol for maintaining multidatabase referential integrity. *Pac. Symp. Biocomput.* 438-445.

Karp, P. (1998a). What we do not know about sequence analysis and sequence databases. *Bioinformatics,* **14,** 753-754.

Karp, P. D., Riley, M., Paley, S. M., Pellegrini-Toole, A. & Krummenacker, M. (1998b). EcoCyc: encyclopedia of *Escherichia coli* genes and metabolism. *Nucl. Acids Res.* **26,** 50-53.

Karp, P. D., Ouzounis, C. & Paley, S. M. (1996b). HinCyc: a knowledge base of the complete genome and metabolic pathways of *H. influenzae. ISMB,* **4,** 116-124.

Lesk, A. M. & Chothia, C. (1984). Mechanisms of domain closure in proteins. *J. Mol. Biol.* **174,** 175-191.

Levitt, M. & Gerstein, M. (1998). A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl Acad. Sci. USA,* **95,** 5913-5920.

Mewes, H. W., Hani, J., Pfeiffer, F. & Frishman, D. (1998). MIPS: a database for protein sequences and complete genomes. *Nucl. Acids Res.* **26,** 33-37.

Moult, J., Hubbard, T., Fidelis, K. & Pedersen, J. T. (1997). Critical assessment of methods of protein structure prediction (CASP): round II. *Proteins: Struct. Funct. Genet.* **1,** 2-6.

Murzin, A., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins for the investigation of sequences and structures. *J. Mol. Biol.* **247,** 536-540.

Myers, E. & Miller, W. (1988). Optimal alignments in linear space. *Comput. Appl. Biosci.* **4,** 11-17.

Needleman, S. B. & Wunsch, C. D. (1971). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48,** 443-453.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. & Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of genes and genomes. *Nucl. Acids Res.* **27,** 29-34.

Park, J., Teichmann, S. A., Hubbard, T. & Chothia, C. (1997). Intermediate sequences increase the detection of homology between sequences. *J. Mol. Biol.* **273,** 349-354.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. & Chothia, C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284,** 1201-1210.

Pearson, W. R. (1996). Effective protein sequence comparison. *Methods Enzymol.* **266,** 227-259.

Pearson, W. R. (1998). Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* **276,** 71-84.

Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA,* **85,** 2444-2448.

Riley, M. (1998a). Systems for categorizing functions of gene products. *Curr. Opin. Struct. Biol.* **8,** 388-392.

Riley, M. (1998b). Genes and proteins of *Escherichia coli* K-12. *Nucl. Acids Res.* **26,** 54.

Riley, M. & Labedan, B. (1996). *E. coli* gene products: physiological functions and common ancestries. In Escherichia coli *and* Salmonella: *Cellular and Molecular Biology* (Neidhardt, F., Curtiss, R., III, Lin, E. C. C., Ingraham, J., Low, K. B., Magasanik, B., Reznikoff, W., Riley, M., Schaechter, M. & Umbarger, H. E., eds), 2nd edit., pp. 2118-2202, ASM Press, Washington, DC.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng.* **12,** 85-94.

Russell, R. B. & Barton, G. J. (1994). Structural features can be unconserved in proteins with similar folds. *J. Mol. Biol.* **244,** 332-350.

Russell, R. B., Saqi, M. A. S., Sayle, R. A., Bates, P. A. & Sternberg, M. J. E. (1997). Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. *J. Mol. Biol.* **269,** 423-439.

Russell, R. B., Sasieni, P. D. & Sternberg, M. J. E. (1998). Supersites within superfolds - binding site similarity in the absence of homology. *J. Mol. Biol.* **282,** 903-918.

Salamov, A. A., Suwa, M., Orengo, C. A. & Swindells, M. B. (1999). Combining sensitive database searches with multiple intermediates to detect distant homologues. *Protein Eng.* **12,** 95-100.

Selkov, E., Jr, Grechkin, Y., Mikhailova, N. & Selkov, E. (1998). MPW: the metabolic pathways database. *Nucl. Acids Res.* **26,** 43-45.

Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147,** 195-198.

Sternberg, M. J. E., Bates, P. A., Kelley, L. A. & MacCallum, R. M. (1999). Progress in protein structure prediction: assessment of CASP3. *Curr. Opin. Struct. Biol.* **9,** 368-373.

Tamames, J., Casari, G., Ouzounis, C. & Valencia, A. (1997). Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.* **44,** 66-73.

Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997). A genomic perspective on protein families. *Science,* **278,** 631-637.

Webb, E. C. (1992). *Enzyme Nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology,* Academic Press, New York.

Wood, T. C. & Pearson, W. R. (1999). Evolution of protein sequences and structures. *J. Mol. Biol.* **291,** 977-995.

Zhang, Z., Schäffer, A. A., Miller, W., Madden, T. L., Lipman, D. J., Koonin, E. V. & Altschul, S. F. (1998). Protein sequence similarity searches using patterns as seeds. *Nucl. Acids Res.* **26,** 3986-3990.

Edited by F. E. Cohen

(Received 2 September 1999; received in revised form 5 January 2000; accepted 6 January 2000)

Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 9193–9197, December 1989
Biochemistry

Isolation and characterization of *ERBB3*, a third member of the *ERBB*/epidermal growth factor receptor family: Evidence for overexpression in a subset of human mammary tumors

(receptor-like kinase/sequence/chromosomal mapping/expression)

MATTHIAS H. KRAUS*, WOLFGANG ISSING*, TORU MIKI*, NICHOLAS C. POPESCU†, AND STUART A. AARONSON*

*Laboratory of Cellular and Molecular Biology and †Laboratory of Biology, National Cancer Institute, Bethesda, MD 20892

Communicated by Leon A. Heppel, August 28, 1989

ABSTRACT A related DNA fragment distinct from the epidermal growth factor receptor and *ERBB2* genes was detected by reduced stringency hybridization of v-*erbB* to normal genomic human DNA. Characterization of the cloned DNA fragment mapped the region of v-*erbB* homology to three exons with closest identity of 64% and 67% to a contiguous region within the tyrosine kinase domains of the epidermal growth factor receptor and ERBB2 proteins, respectively. cDNA cloning revealed a predicted 148-kDa transmembrane polypeptide with structural features identifying it as a member of the *ERBB* gene family, prompting us to designate the gene as *ERBB3*. It was mapped to human chromosome 12q13 and was shown to be expressed as a 6.2-kilobase transcript in a variety of normal tissues of epithelial origin. Markedly elevated ERBB3 mRNA levels were demonstrated in certain human mammary tumor cell lines. These findings suggest that increased *ERBB3* expression, as in the case of epidermal growth factor receptor and *ERBB2*, may play a role in some human malignancies.

Protooncogenes encoding growth factor receptors constitute several distinct families with close overall structural homology. The highest degree of homology is observed in their catalytic domains, essential for the intrinsic tyrosine kinase activity of these proteins (1). Examples of such families include genes encoding epidermal growth factor receptor (EGF-R) and ERBB2, the colony-stimulating factor 1/platelet-derived growth factor receptors, insulin/insulin-like growth factor 1 receptors, and EPH/ELK (2–12). Growth factor receptors in several of these families play critical roles in regulation of normal growth and development. Some of these molecules have been implicated in the neoplastic process as well. In particular, both the EGF-R gene and *ERBB2* have been shown to be activated as oncogenes by mechanisms involving overexpression or mutations that constitutively activate the catalytic activity of their encoded proteins (13–16). Thus, we undertook the present studies in an effort to identify and isolate additional members of the *ERBB* protooncogene family.

MATERIALS AND METHODS

Human Cells. Mammary epithelial cells AB589 (17) and immortalized keratinocytes RHEK (18) were provided by M. Stampfer (Lawrence Berkeley Laboratory) and J. Rhim, respectively. Normal human epidermal melanocytes (NHEM) and keratinocytes (NHEK) were obtained from Clonetics (San Diego, CA). Sources for human embryo fibroblasts (19) or mammary tumor cell lines (20) have been described.

DNA and RNA Hybridization. High-stringency hybridization was conducted as described (20). Reduced-stringency hybridization of DNA was carried out in 30% (vol/vol) formamide followed by washes in 0.6× SSC, whereas intermediate stringency was achieved by hybridization in 40% (vol/vol) formamide and washing in 0.25× SSC.

Molecular Cloning. An oligo(dT)-primed human placenta cDNA library was obtained from Clontech. The oligo(dT) primed MCF-7 cDNA library was constructed in λpCEV9 (21). After plaque purification, phage DNA inserts were subcloned into pUC-based plasmid vectors for further characterization.

Nucleotide and Amino Acid Sequence Analysis. The nucleotide sequence was determined for both DNA strands by the dideoxy chain-termination method (22) using supercoiled plasmid DNA as template.‡ Amino acid sequence comparison was performed with the alignment program by Pearson and Lipman (23). Hydrophobic and hydrophilic regions in the predicted protein were identified according to Kyte and Doolittle (24).

RESULTS

Identification of a Third Member of the *ERBB* Protooncogene Family. In an effort to detect novel *ERBB*-related genes, human genomic DNA was cleaved with a variety of restriction endonucleases and subjected to Southern blot analysis with v-*erbB* as probe. Under reduced stringency hybridization, four *Sac* I restriction fragments were detected. Two were identified as EGF-R gene fragments by their amplification in MDA-MB468 cells (Fig. 1*A*, lanes 1 and 2) known to contain EGF-R gene amplification and one as an *ERBB2*-specific gene fragment due to its increased signal intensity in *ERBB2*-amplified SK-BR-3 cells (Fig. 1*A*, lanes 1 and 3). However, a single 9-kbp *Sac* I fragment exhibited equal signal intensities in normal human thymus, A431, and SK-BR-3 DNA (Fig. 1*A*). When the hybridization stringency was raised by 7°C, this fragment did not hybridize, whereas EGF-R and *ERBB2*-specific restriction fragments were still detected with v-*erbB* as a probe (Fig. 1*B*). Taken together, these findings suggested the specific detection of another v-*erbB*-related DNA sequence within the 9-kbp *Sac* I fragment.

For further characterization we prepared a normal human genomic library from *Sac* I-cleaved thymus DNA enriched for 8- to 12-kbp fragments. Ten recombinant clones detected by v-*erbB* under reduced stringency conditions did not hybridize with human EGF-R or ERBB2 cDNA probes at high stringency. As shown in the restriction map of a repre-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: EGF, epidermal growth factor.
‡The *ERBB3* nucleotide sequence has been deposited in the GenBank data base (accession no. M29366).

sentative clone with a 9-kbp insert, the region of v-*erbB* homology was localized by hybridization analysis to a 1.5-kbp segment spanning from the *Eco*RI to the downstream *Pst* I site. Nucleotide sequence analysis revealed that this region contained three open reading frames bordered by splice junction consensus sequences (Fig. 2). The predicted amino acid sequence of these three open reading frames revealed the highest identity scores of 64–67% to three regions that are continuous in the tyrosine kinase domains of v-*erbB*, as well as human EGF-R and ERBB2 proteins. Furthermore, all splice junctions of the three characterized exons in the gene were conserved with *ERBB2*. Amino acid sequence homology to other known tyrosine kinases was significantly lower, ranging from 39–46%.

A single 6.2-kb-specific mRNA was identified by Northern (RNA) blot analysis of human epithelial cells by using the 150-base pair (bp) *Spe* I-*Acc* I exon-containing fragment as probe (Fig. 2). Under the stringent hybridization conditions used, this probe detected neither the 5-kb ERBB2 mRNA nor the 6- and 10-kb EGF-R mRNAs (data not shown). All of these findings suggested that we had identified an additional functional member of the *ERBB* protooncogene family, which we tentatively designated as *ERBB3*.

Close Structural Similarity of the Predicted ERBB3 Protein with Other *ERBB* Family Members. In an effort to characterize the entire *ERBB3* coding sequence, overlapping cDNA clones were isolated from oligo(dT)-primed cDNA libraries from sources with known *ERBB3* expression, utilizing gene-specific genomic exons or cDNA fragments as probes. The clones were initially characterized by restriction analysis and hybridization to the mRNA and were subsequently subjected to nucleotide sequence analysis. The clones pE3-8, pE3-9, pE3-11, and pE3-16 contained identical 3′ ends terminating in a poly(A) stretch (Fig. 2).

The complete coding sequence of *ERBB3* was contained within a single long open reading frame of 4080 nucleotides extending from position 46 to an in-frame termination codon at position 4126. The most upstream ATG codon at position 100 was the likely initiation codon, as it was preceded by an in-frame stop codon at nucleotide position 43 and fulfilled Kozak's criteria (26) for an authentic initiation codon. The open reading frame comprised 1342 codons, predicting a 148-kDa polypeptide. As shown in Fig. 3, the deduced amino acid sequence of the ERBB3 polypeptide predicted a transmembrane receptor tyrosine kinase most closely related to EGF-R and ERBB2. A hydrophobic signal sequence of ERBB3 was predicted to comprise the 19 amino-terminal amino acid residues. Cleavage of this signal sequence between Gly-19 and Ser-20 would generate a processed polypeptide of 1323 amino acids with an estimated molecular mass of 145 kDa. A single hydrophobic membrane-spanning domain encompassing 21 amino acids was identified (Fig. 3).

The putative ERBB3 ligand-binding domain was 43% and 45% identical in amino acid residues with the predicted ERBB2 and EGF-R protein, respectively. Within the extracellular domain, all 50 cysteine residues of the processed ERBB3 polypeptide were conserved and similarly spaced when compared with the EGF-R and ERBB2. Forty-seven cysteine residues were organized in two clusters containing 22 and 25 cysteines, respectively, a structural hallmark of this tyrosine kinase receptor subfamily (2–4). Ten potential N-linked glycosylation sites were localized within the ERBB3 extracellular domain. In comparison with the EGF-R and ERBB2 proteins, five and two of these glycosylation sites were conserved, respectively. Among these, the site proximal to the transmembrane domain was conserved among all three proteins (Fig. 3).

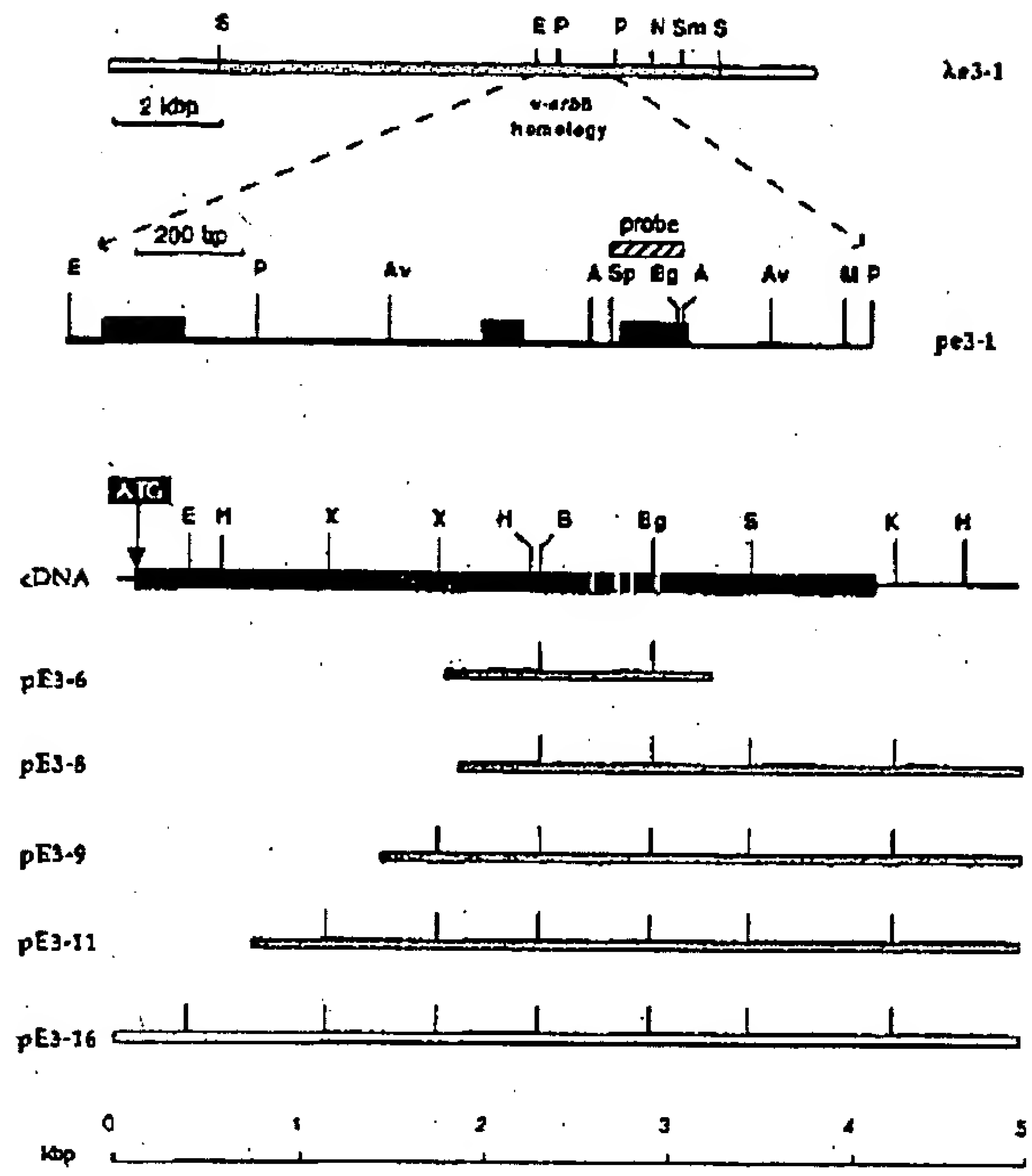


FIG. 2. Genomic and cDNA cloning of *ERBB3*. The region of v-*erbB* homology within the genomic 9-kbp *Sac* I insert of λe3-1 was subcloned into pUC (pe3-1) and subjected to nucleotide sequence analysis. The three predicted exons are depicted as solid boxes. ERBB3 cDNA clones were isolated from normal human placenta (shaded bars) and MCF-7 (open bar) oligo(dT)-primed libraries. The entire nucleotide sequence was determined for both strands on ERBB3 cDNA from normal human placenta and upstream of the 5′ *Xho* I site on pE3-16. The coding sequence is shown as a solid bar, and splice junctions of the three characterized genomic exons are indicated by vertical white lines. Solid lines in the cDNA map represent untranslated sequences. Restriction sites: A, *Acc* I; Av, *Ava* I; B, *Bam*HI; Bg, *Bgl* II; E, *Eco*RI; H, *Hin*dIII; K, *Kpn* I; M, *Mst* II; P, *Pst* I; S, *Sac* I; Sm, *Sma* I; and Sp, *Spe* I.

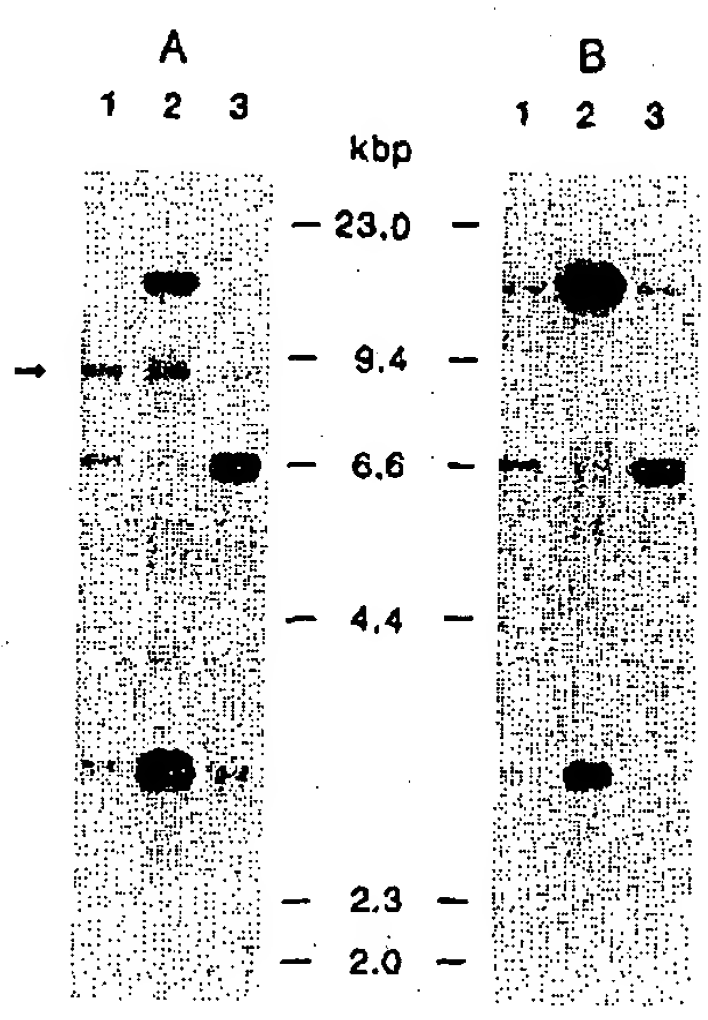


FIG. 1. Detection of v-*erbB*-related gene fragments in normal human thymus (lane 1), MDA-MB468 (lane 2), SK-BR-3 (lane 3). DNAs were restricted with *Sac* I and hybridized with a v-*erbB*-specific probe spanning from the upstream *Bam*HI to the *Eco*RI site in avian erythroblastosis proviral DNA (25). Hybridization was conducted at reduced (*A*) or intermediate (*B*) stringency conditions. The arrow denotes a 9-kbp *ERBB*-related restriction fragment distinct from those of EGF-R and *ERBB2*.

Within the cytoplasmic domain, a core of 277 amino acids from position 702–978 revealed the most extensive homology with the tyrosine kinase domains of EGF-R and ERBB2. In this region 60 or 62% of amino acid residues were identical and 90 or 89% were conserved, respectively. This stretch of amino acid homology coincides with the minimal catalytic domain of tyrosine kinases (1). There was significantly lower homology with other tyrosine kinases (Fig. 3). The consensus sequence for an ATP-binding site Gly-Xaa-Gly-Xaa-Xaa-Gly (1) at amino acid position 716–721 as well as a lysine residue located 21 amino acid residues farther carboxyl-terminal were conserved between the three ERBB-related receptors. Taken together, these findings defined the region between amino acid position 702 and 978 as the putative catalytic domain of the ERBB3 protein (Fig. 3).

The most divergent region of ERBB3 compared with either EGF-R or ERBB2 was its carboxyl terminus, comprising 364 amino acids. Tyrosine residues at positions 1197, 1199, and 1262 matched closest with the consensus sequence for putative phosphorylation sites (28). The peptide one-letter sequence YEYMN, encompassing Tyr-1197 and Tyr-1199, was repeated at positions 1260–1264 and was at both locations surrounded by charged residues. These observations render Tyr-1197, Tyr-1199, and Tyr-1262 likely candidates for autophosphorylation sites of the ERBB3 protein.

Chromosomal Mapping of Human *ERBB3*. We determined the chromosomal location of the *ERBB3* gene by *in situ* hybridization (29) with an ^{3}H-labeled plasmid containing the *ERBB3* amino-terminal coding sequence. A total of 110 human chromosome spreads were examined prior and subsequent to G banding for identification of individual chromosomes. One hundred forty-two grains were localized on a 400-band ideogram. We observed specific labeling of chromosome 12, where 38 out of 51 grains were localized to band q13 (data not shown). Thus, the genomic locus of *ERBB3* was assigned to 12q13. In this region of chromosome 12, several genes have previously been mapped including the melanoma-associated antigen ME491 (30), histone genes (31) and the gene for lactalbumin (32). In addition, two protooncogenes, *INT1* (33) and *GLI* (34), are located in close proximity to *ERBB3*.

***ERBB3* Expression in Normal and Malignant Human Cells.** To investigate its pattern of expression, we surveyed a number of human tissues for the ERBB3 transcript. The 6.2-kb ERBB3-specific mRNA was observed in term placenta, postnatal skin, stomach, lung, kidney, and brain, but it was not detectable in skin fibroblasts, skeletal muscle, or lymphoid cells (data not shown). Among the fetal tissues analyzed, the ERBB3 transcript was expressed in liver, kidney, and brain but not in fetal heart or embryonic lung fibroblasts. These

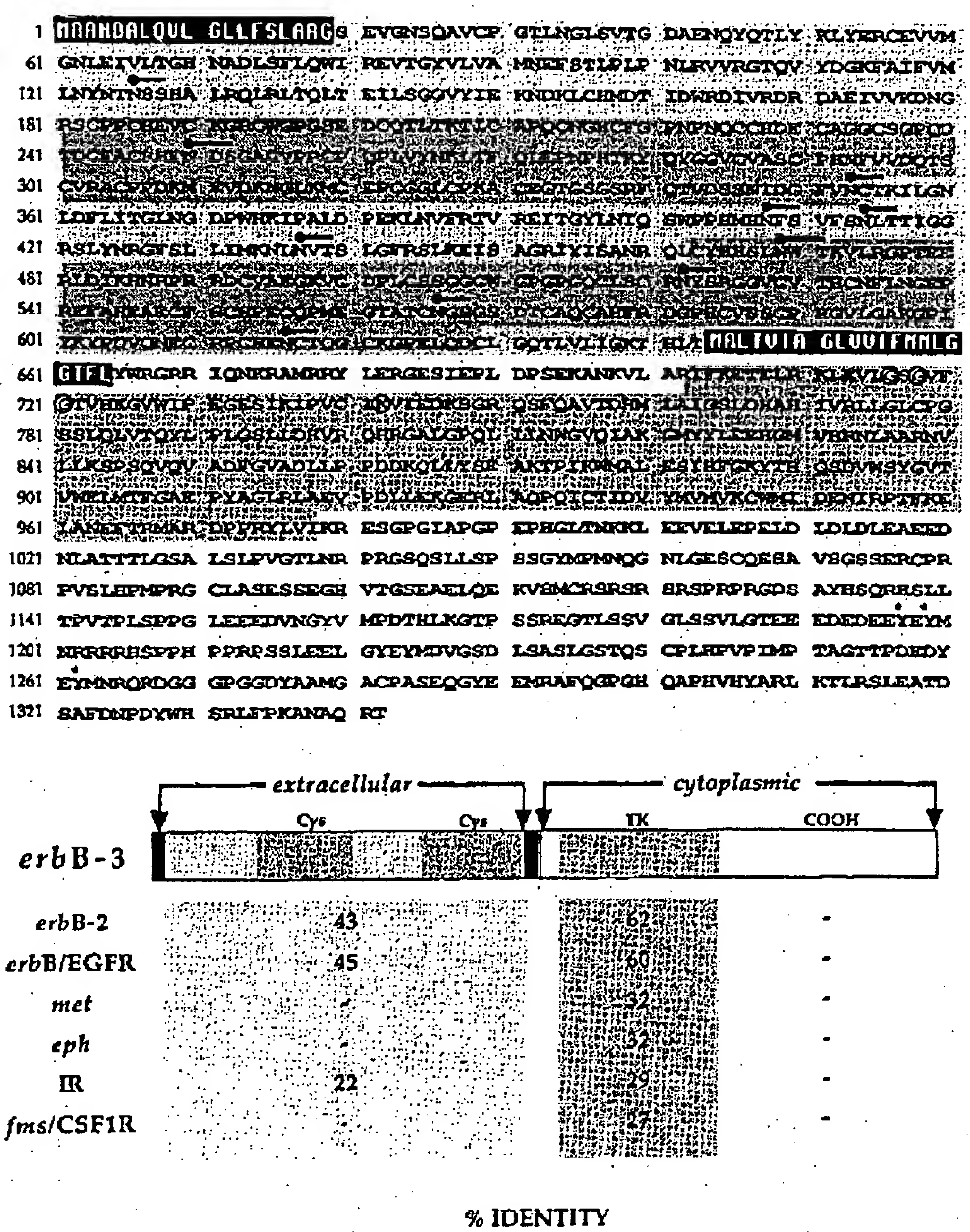


FIG. 3. (*Upper*) Predicted amino acid sequence of the ERBB3 polypeptide (as deduced from the GenBank sequence M29366) and comparison with other receptor-like tyrosine kinases. The amino acid sequence is shown in single-letter code and is numbered at left. The putative extracellular domain (light shading) extends between the predicted signal sequence (solid box) at the amino terminus and a single hydrophobic transmembrane region (solid box) within the polypeptide. The putative ATP-binding site at the amino terminus of the TK domain is circled. Potential autophosphorylation sites within the carboxyl-terminal domain (COOH) are indicated by asterisks. Potential N-linked glycosylation sites (●—) are marked above the amino acid sequence. (*Lower*) The two cysteine clusters (Cys) in the extracellular domain and the predicted tyrosine kinase domain (TK) within the cytoplasmic portion of the polypeptide are outlined by dark shading. The percentage of amino acid homology of ERBB3 in individual domains with ERBB2 (4), EGF-R (2), MET (27), EPH (11), insulin receptor [IR (9)], and FMS (5) is listed below. Less than 16% identity is denoted by -.

observations indicated the preferential expression of the ERBB3 transcript in epithelial tissues and brain.

We also investigated *ERBB3* expression in individual cell populations in comparison to EGF-R and ERBB2 transcripts. As shown in Table 1, mRNA levels of each were relatively high in keratinocytes and low but similar in cells derived from glandular epithelium. These findings are consistent with growth regulatory roles of all three receptor-like molecules in squamous and glandular epithelium. Whereas ERBB2 and EGF-R transcripts were also readily seen in normal fibroblasts, the same cells lacked detectable ERBB3 mRNA. In contrast, normal human melanocytes, which expressed both ERBB3 and ERBB2 at levels comparable with human keratinocytes, lacked detectable EGF-R transcripts. Thus, the expression patterns of these receptor-like molecules were different in specialized cell populations derived from epidermal tissues.

The ERBB3 transcript was detected in 36 of 38 carcinomas and 2 of 12 sarcomas, whereas 7 tumor cell lines of hematopoetic origin lacked measurable ERBB3 mRNA. Markedly elevated levels of a normal-sized transcript were observed in 6 of 17 tumor cell lines derived from human mammary carcinomas. By Southern blot analysis, neither gross gene rearrangement nor amplification was detected in the cell lines (data not shown). Fig. 4*A* shows the results of Northern blot analysis with control AB589 nonmalignant human mammary epithelial cells (lane 1) and two representative human mammary tumor lines, MDA-MB415 (lane 2) and MDA-MB453 (lane 3). Hybridization of the same filter with a human β actin probe (Fig. 4*B*) verified levels of mRNA in each lane. Densitometric scanning indicated that the ERBB3 transcript in each tumor cell line was elevated more than 100-fold above that of the control cell line. Thus, overexpression of this third member of the *ERBB* family, as for the EGF-R and *ERBB2* genes, may play an important role in some human malignancies.

DISCUSSION

In the present report, we describe the identification of a third member of the *ERBB*/EGF receptor family of membrane-spanning tyrosine kinases and the cloning of its full-length coding sequence. This gene, designated *ERBB3*, encodes a predicted protein with striking structural similarities to other members of this family. These features include overall size, extracellular domain with two signature cysteine clusters, and its uninterrupted tyrosine kinase domain, exhibiting with 81% and 83% significantly greater overall similarities to EGF-R and ERBB2 products than to any other tyrosine kinase. The structural relatedness of its extracellular domain with that of the EGF-R raises the possibility that one or more of an increasing number of EGF-like ligands (36) may interact with the ERBB3 product.

Distinct regions within the predicted *ERBB3* coding sequence revealed relatively higher degrees of divergence. For example, its carboxyl-terminal domain failed to exhibit significant colinear identity scores with either ERBB2 or EGF-R. Within the tyrosine kinase domain, which represents the most conserved region of the predicted ERBB3 protein, a short stretch of 29 amino acids carboxyl-terminal to the ATP-binding site differed from regions of the predicted ERBB2 and EGF-R coding sequence in 28 and 25 positions, respectively. Such regions of higher divergence in their cytoplasmic domains may confer functional specificity to these closely related receptor-like molecules.

Chromosomal mapping localized *ERBB3* to human chromosome 12q11-13, whereas the related EGF-R and *ERBB2* genes are located on chromosomes 7p12-13 (37) and 17p12-21.3 (3, 29), respectively. Thus, each appears to localize to regions containing different respective homeobox (38, 39) and collagen gene (40) loci. Keratin type I and type II genes also map to regions of 12 and 17 (41, 42), consistent with localization of *ERBB3* and *ERBB2*, respectively.

Recent studies in *Drosophila* have emphasized how critical and multifunctional are developmental processes mediated by ligand–receptor interactions. An increasing number of *Drosophila* mutants with often varying phenotypes have now been identified as being due to lesions in genes encoding such proteins (43, 44) including the *Drosophila* EGF-R homologue, designated DER. It is not yet known whether DER is the Drosophila counterpart of all three mammalian *ERBB* genes. If so, functions assigned to DER may eventually be associated with one or more of the divergent mammalian *ERBB* genes as well as other functions that have evolved in more complex mammalian organisms.

There is evidence for autocrine (45, 46) as well as paracrine (19, 47) effectors of normal cell proliferation. However, the inherent transforming potential of autocrine growth factors (48, 49) suggests that growth factors most commonly act on their target cell populations by a paracrine route. Our survey of *ERBB3* gene expression indicated its normal expression in cells of epithelial and neuroectodermal derivation. Comparative analysis of the three *ERBB* receptor-like genes in different cell types of epidermal tissue revealed that keratinocytes expressed all three genes. In contrast, melanocytes and stromal fibroblasts specifically lacked EGF-R and

Table 1. Normal expression pattern of human *ERBB* gene family members

Source	Relative transcript levels ERBB3	ERBB2	EGF-R
Embryonic fibroblast (M426)	−	+	+
Skin fibroblast (501T)	−	+	+
Immortal keratinocyte (RHEK)	++	++	++
Primary keratinocyte (NHEK)	+	+	++
Glandular epithelium (AB589)	(+)	(+)	(+)
Melanocyte (NHEM)	++	++	−

Replicate Northern blots were hybridized with equal probe counts of similar specific activity for ERBB3, ERBB2, and EGF receptor, respectively. Relative signal intensities were semiquantitatively estimated: −, not detectable; (+), weakly positive; +, positive; ++, strongly positive.

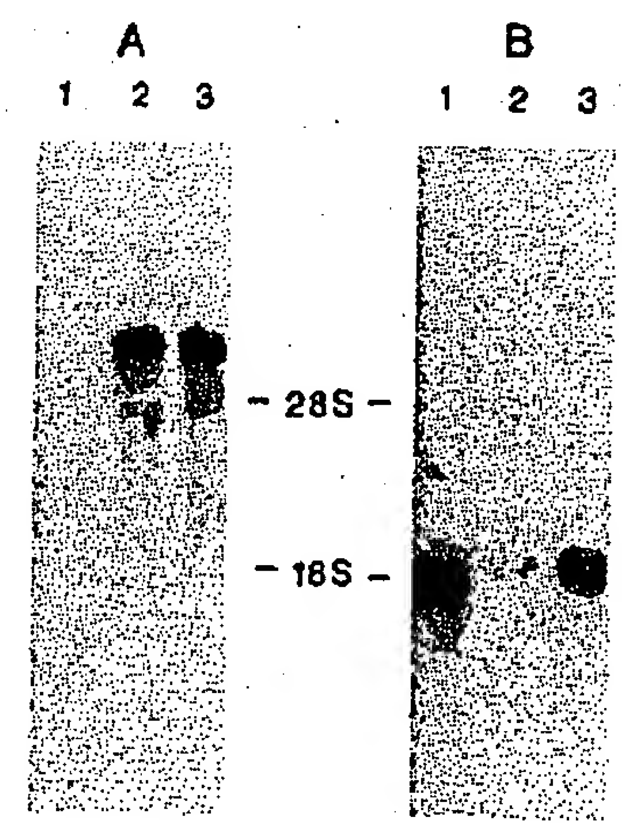


FIG. 4. Elevated ERBB3 transcript levels in human mammary tumor cell lines. A Northern blot containing 10 μg of total cellular RNA from AB589 mammary epithelial cells (lane 1), as well as MDA-MB415 (lane 2) and MDA-MB453 (lane 3) mammary tumor cell lines was hybridized with an ERBB3 cDNA probe (*A*). After signal decay the same blot was rehybridized with a human β actin cDNA probe (35).

ERBB3 transcripts, respectively. Thus, melanocytes and stromal fibroblasts may be sources of paracrine growth factors for EGF-R and ERBB3 products, respectively, that are expressed by the other cell types residing in close proximity in epidermal tissues.

To date, both *ERBB*/EGFR and *ERBB2* have been causally implicated in human malignancy. EGF-R gene amplification and/or overexpression in tumors has been demonstrated in squamous cell carcinomas and glioblastomas (50). *ERBB2* amplification and/or overexpression has been observed in human breast and ovarian carcinomas (51, 52), and *ERBB2* overexpression has been reported to be an important prognostic indicator of particularly aggressive tumors (52). Thus, our present findings that the ERBB3 transcript is overexpressed in a significant fraction of human mammary tumor cell lines raises the possibility that this new member of the ERBB/EGF receptor family may also play an important role in some human malignancies.

We thank Drs. D. Ron, G. Kruh, and P. Finch for providing some of the cellular RNA samples used in the initial *ERBB3* expression analysis.

1. Hanks, S. K., Quinn, A. M. & Hunter, T. (1988) *Science* 241, 42–52.
2. Ullrich, A., Coussens, L., Hayflick, J. S., Dull, T. J., Gray, A., Tam, A. W., Lee, J., Yarden, Y., Libermann, T. A., Schlessinger, J., Downward, J., Mayes, E. L. V., Whittle, N., Waterfield, M. D. & Seeburg, P. H. (1984) *Nature (London)* 309, 418–425.
3. Coussens, L., Yang-Feng, T. L., Liao, Y. C., Chen, E., Gray, A., McGrath, J., Seeburg, P. H., Libermann, T. A., Schlessinger, J., Francke, U., Levinson, A. & Ullrich, A. (1985) *Science* 230, 1132–1139.
4. Yamamoto, T., Ikawa, S., Akiyama, T., Semba, K., Nomura, N., Miyajima, N., Saito, T. & Toyoshima, K. (1986) *Nature (London)* 319, 230–234.
5. Coussens, L., van Beveren, C., Smith, D., Chen, E., Mitchell, R. L., Isacke, C. M., Verma, I. M. & Ullrich, A. (1986) *Nature (London)* 320, 277–280.
6. Sherr, C. J., Rettenmier, C. W., Sacca, R., Roussel, M. F., Look, A. T. & Stanley, E. R. (1985) *Cell* 41, 665–676.
7. Yarden, Y., Escobedo, J. A., Kuang, W.-J., Yang-Feng, T. L., Daniel, T. O., Tremble, P. M., Chen, E. Y., Ando, M. E., Harkins, R. N., Francke, U., Fried, V. A., Ullrich, A. & Williams, L. T. (1986) *Nature (London)* 323, 226–232.
8. Matsui, T., Heidaran, M., Miki, T., Popescu, N., La Rochelle, W., Kraus, M., Pierce, J. & Aaronson, S. (1989) *Science* 243, 800–804.
9. Ullrich, A., Bell, J. R., Chen, E. Y., Herrera, R., Petruzzelli, L. M., Dull, T. J., Gray, A., Coussens, L., Liao, Y.-C., Tsubokawa, M., Mason, A., Seeburg, P. H., Grunfeld, C., Rosen, O. M. & Ramachandran, J. (1985) *Nature (London)* 313, 756–761.
10. Ullrich, A., Gray, A., Tam, W., Yang-Feng, T., Tsubokawa, M., Collins, C., Henzel, W., Le Bon, T., Kathuria, S., Chen, E., Jacobs, S., Francke, U., Ramachandran, J. & Fujita-Yamaguchi, Y. (1986) *EMBO J.* 5, 2503–2512.
11. Hirai, H., Maru, Y., Hagiwara, K., Nishida, J. & Takaku, F. (1987) *Science* 238, 1717–1720.
12. Letwin, K., Yee, S.-P. & Pawson, T. (1988) *Oncogene* 3, 621–627.
13. Bargmann, C. I., Hung, M. C. & Weinberg, R. A. (1986) *Cell* 45, 649–657.
14. Di Fiore, P. P., Pierce, J. H., Kraus, M. H., Segatto, O., King, C. R. & Aaronson, S. A. (1987) *Science* 237, 178–182.
15. Di Fiore, P. P., Pierce, J. H., Fleming, T. P., Hazan, R., Ullrich, A., King, C. R., Schlessinger, J. & Aaronson, S. A. (1987) *Cell* 51, 1063–1070.
16. Velu, T. J., Beguinot, L., Vass, W. C., Willingham, M. C., Merlino, G. T., Pastan, I. & Lowy, D. R. (1987) *Science* 238, 1408–1410.
17. Walen, K. H. & Stampfer, M. R. (1989) *Cancer Genet. Cytogenet.* 37, 249–261.
18. Rhim, J. S., Jay, G., Arnstein, P., Price, F. M., Sanford, K. K. & Aaronson, S. A. (1985) *Science* 227, 1250–1252.
19. Rubin, J. S., Osada, H., Finch, P. W., Taylor, W. G. & Rudikoff, S. & Aaronson, S. A. (1989) *Proc. Natl. Acad. Sci. USA* 86, 802–806.
20. Kraus, M. H., Popescu, N. C., Amsbaugh, S. C. & King, C. R. (1987) *EMBO J.* 6, 605–610.
21. Miki, T., Matsui, T., Heidaran, M. A. & Aaronson, S. A. (1989) *Gene*, in press.
22. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* 74, 5463–5467.
23. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* 85, 2444–2448.
24. Kyte, J. & Doolittle, R. F. (1982) *J. Mol. Biol.* 157, 105–132.
25. Vennstrom, B., Fanshier, L., Moscovici, C. & Bishop, J. M. (1980) *J. Virol.* 36, 575–585.
26. Kozak, M. (1987) *Nucleic Acids Res.* 15, 8125–8148.
27. Park, M., Dean, M., Kaul, K., Braun, M. J., Gonda, M. A. & van de Woude, G. (1987) *Proc. Natl. Acad. Sci. USA* 84, 6379–6383.
28. Smart, J. E., Oppermann, H., Czernilofsky, A. P., Purchio, A. F., Erikson, R. L. & Bishop, J. M. (1981) *Proc. Natl. Acad. Sci. USA* 78, 6013–6017.
29. Popescu, N. C., King, C. R. & Kraus, M. H. (1989) *Genomics* 4, 362–366.
30. Hotta, H., Ross, A. H., Huebner, K., Isobe, M., Wendeborn, S., Chao, M. V., Ricciardi, R. P., Tsujimoto, Y., Croce, C. M. & Koprowski, H. (1988) *Cancer Res.* 48, 2955–2962.
31. Tripputi, P., Emanuel, B. S., Croce, C. M., Green, L. G., Stein, G. S. & Stein, J. L. (1986) *Proc. Natl. Acad. Sci. USA* 83, 3185–3188.
32. Davies, M. S., West, L. F., Davies, M. B., Povey, S. & Craig, R. K. (1987) *Ann. Hum. Genet.* 51, 183–188.
33. Turc-Carel, C., Pietrzak, E., Kakati, S., Kinniburgh, A. J. & Sandberg, A. A. (1987) *Oncogene Res.* 1, 397–405.
34. Kinsler, K. W., Bigner, S. H., Bigner, D. D., Trent, J. M., Law, M. L., O'Brien, S. J., Wong, A. J. & Vogelstein, B. (1987) *Science* 236, 70–73.
35. Gunning, P., Ponte, P., Okayama, H., Engel, J., Blau, H. & Kedes, L. (1983) *Mol. Cell. Biol.* 3, 787–795.
36. Shoyab, M., Plowman, G. D., McDonald, V. L., Bradley, J. G. & Todaro, G. J. (1989) *Science* 243, 1074–1076.
37. Spurr, N. K., Solomon, E., Jansson, M., Sheer, D., Goodfellow, P. N., Bodmer, W. F. & Vennstrom, B. (1984) *EMBO J.* 3, 159–163.
38. Rabin, M., Ferguson-Smith, A., Hart, C. P. & Ruddle, F. H. (1986) *Proc. Natl. Acad. Sci. USA* 83, 9104–9108.
39. Cannizzaro, L. A., Huebner, K., Griffen, C. A., Simeone, A., Boncinelli, E. & Croce, C. M. (1987) *Am. J. Hum. Genet.* 41, 1–15.
40. Retief, E., Parker, M. I. & Retief, A. E. (1985) *Hum. Genet.* 69, 304–308.
41. Rosenberg, M., RayChaudury, A., Shows, T. B., Le Beau, M. M. & Fuchs, E. (1988) *Mol. Cell. Biol.* 8, 722–736.
42. Popescu, N. C., Bowden, P. & DiPaolo, J. A. (1989) *Hum. Genet.* 82, 109–112.
43. Schejter, E. D. & Shilo, B. Z. (1989) *Cell* 56, 1093–1104.
44. Price, J. V., Clifford, R. J. & Schüpbach, T. (1989) *Cell* 56, 1085–1092.
45. Sporn, M. B. & Todaro, G. J. (1980) *N. Engl. J. Med.* 303, 878–880.
46. Kaplan, P. L., Anderson, M. & Ozanne, B. (1982) *Proc. Natl. Acad. Sci. USA* 79, 485–489.
47. Weiner, H. L. & Swain, J. L. (1989) *Proc. Natl. Acad. Sci. USA* 86, 2683–2687.
48. Doolittle, R. F., Hunkapiller, M. W., Hood, L. E., Devare, S. G., Robbins, K. C., Aaronson, S. A. & Antoniades, H. N. (1983) *Science* 221, 275–277.
49. Waterfield, M. D., Scrace, G. T., Whittle, N., Stroobant, P., Johnsson, A., Wasteson, A., Westermark, B., Heldin, C. H., Huang, J. S. & Deuel, T. F. (1983) *Nature (London)* 304, 35–39.
50. Libermann, T. A., Nusbaum, H. R., Razon, N., Kris, R., Lax, I., Soreq, H., Whittle, N., Waterfield, M. D., Ullrich, A. & Schlessinger, J. (1985) *Nature (London)* 313, 144–147.
51. King, C. R., Kraus, M. H. & Aaronson, S. A. (1985) *Science* 229, 974–976.
52. Slamon, D. J., Godolphin, W., Jones, L. A., Holt, J. A., Wong, S. G., Keith, D. E., Levin, W. J., Stuart, S. G., Udove, J., Ullrich, A. & Press, M. F. (1989) *Science* 244, 707–712.

THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 270, No. 28, Issue of July 14, pp. 16510–16513, 1995
Printed in U.S.A.

Communication

Characterization of a Stat-like DNA Binding Activity in *Drosophila melanogaster**

(Received for publication, May 17, 1995, and in revised form, May 26, 1995)

Sharon M. Sweitzer‡§, Soledad Calvo¶, Matthias H. Kraus¶, David S. Finbloom‡, and Andrew C. Larner‡‖

From the ‡Division of Cytokine Biology, Food and Drug Administration, Bethesda, Maryland 20892 and ¶Laboratory of Cellular and Molecular Biology, NCI, National Institutes of Health, Bethesda, Maryland 20892

The cytokine signaling pathways that activate the Janus family of tyrosine kinases (Jaks) and the "signal transducers and activators of transcription" (Stats) have been well characterized in mammalian systems. Work shown here provides evidence that an analogous signaling pathway exists in *Drosophila melanogaster*. Because many of the ligand-receptor pairs in *Drosophila* have not been fully characterized, it was necessary to bypass the receptor stimulation event that normally triggers intracellular Jak/Stat activation. This was done by treating *Drosophila* Schneider 2 cells with vanadate/peroxide, which has been shown to closely mimic some signaling events triggered by interferon γ, including the activation of Jak1, Jak2, and the Stat1α protein. Evidence presented here demonstrates that vanadate/peroxide can induce a γ response region binding complex in *Drosophila* Schneider 2 cells. This complex contains two phosphoproteins of 100 and 150 kDa, respectively, and shares many features with the vanadate/peroxide-stimulated binding complex in the mammalian system. Southern blot analysis of genomic DNA using the *src* homology domain 2 (SH2) of Stat1α confirms the presence of a related gene in the *Drosophila* genome.

The Jak/Stat[1] pathway is activated in mammalian cell systems by treatment of cells with a number of different cytokines and growth factors (1, 2). The Stat family of transcription activators has several common structural features, including conserved SH2 domains. The current model of Stat activation is that a tyrosine in the carboxyl terminus of the Stat protein is phosphorylated and acts as an SH2 binding site upon cytokine stimulation (3, 4). The Stat can then form dimers, either with itself or with another member of the Stat family. Dimerization of the Stats is necessary for DNA binding and, ultimately, induction of transcription. Activation of a transcription complex containing phosphorylated Stats can also be induced by treating cells with a combination of sodium orthovanadate and hydrogen peroxide in the absence of cytokines or growth factors (5).

There is growing evidence that a Jak/Stat pathway may exist in *Drosophila*. First, *hopscotch*, a *Drosophila* Jak kinase homologue has been cloned and characterized (6). *hopscotch* is a maternal transcript, expressed in embryonic stripes. Mutations in this gene locus cause abnormalities in the expression patterns of some of the pair-rule and segment-polarity genes. This implicates the Jak kinase being involved in signal transduction pathways controlling segmentation of the developing fly.

Second, a number of growth factor receptor homologs have been cloned from *Drosophila*. *torso* encodes a receptor tyrosine kinase that has a cytoplasmic domain similar to that of the mammalian platelet-derived growth factor receptor and is involved in embryonic pattern formation (7, 8). *der*, the epidermal growth factor receptor homolog, has been cloned and genetically characterized (9). Mutations in the *der* locus cause pleiotropic effects, implicating it in the development of a wide range of tissues (10). Treatment of mammalian cells with either platelet-derived growth factor or epidermal growth factor has been shown to activate Jak/Stat pathways, suggesting that Torso or Der *may activate a Jak/Stat pathway in Drosophila* as well.

Using the ligand-independent activation by vanadate/peroxide treatment, we have attempted to identify a Stat-like activity in *Drosophila* Schneider 2 cells. The specific GRR binding activity, which was seen after cell treatment, was dependent upon tyrosine phosphorylation. This binding complex contained phosphoproteins of 100 and 150 kDa. Detection of Stat1α-related sequences in the *Drosophila* genome also suggested that a Stat homolog exists in *Drosophila*.

MATERIALS AND METHODS

Cells and Reagents—Schneider 2 cells (ATCC CRL 1963) were maintained in Schneider's *Drosophila* medium (Life Technologies, Inc.), 10% fetal bovine serum (Quality Biologics). Sodium orthovanadate and hydrogen peroxide were purchased from Sigma. SDS-polyacrylamide gel electrophoresis was performed on the Novex system. 4G10 antibody was purchased from Upstate Biotechnology Inc. All other reagents were purchased from commercial sources unless otherwise noted.

Schneider 2 Cell Treatment—A solution of 50 mM sodium orthovanadate, 500 mM hydrogen peroxide made in Schneider medium was incubated at 24 °C for 5 min. This solution was added to exponentially growing Schneider 2 cells to a final concentration of 100 μM sodium orthovanadate, 1 mM hydrogen peroxide, and cells were incubated for the indicated times at 24 °C. Cells were washed two times in phosphate-buffered saline. Whole cell lysates were prepared by solubilizing cells for 10 min, on ice, with intermittent vortexing, in a buffer containing 20 mM HEPES, pH 7.0, 10 mM KCl, 1 mM $MgCl_2$, 20% glycerol, 0.1% Nonidet P-40, and 1% Triton X-100. Particulate matter was separated from soluble material by centrifugation at 18,000 × *g* for 5 min. Where indicated, cells were pretreated with 500 nM staurosporin or 30 μg/ml genistein for 30 min. Vanadate/peroxide solution (described above) was added to cells, and the incubation continued for an additional 60 min.

Electrophoretic Mobility Shift Assays—10 μg of soluble protein was diluted in binding buffer to a final concentration of 10 mM Tris, pH 7.4, 5 mM $MgCl_2$, 100 mM KCl, 1 mM dithiothreitol, 50% glycerol, 0.03% Nonidet P-40, and 0.1 mg/ml poly(dI·dC). Double-stranded probe was labeled with [γ-^{32}P]ATP using polynucleotide kinase. 1 ng of probe was added to the extract. A 10- or 50-fold excess of unlabeled competitor oligonucleotide was added as indicated. For sequences of oligonucleotides used, see Fig. 2*D*. Binding reactions proceeded for 5 min at room

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

§ Supported by a National Research Council-Food and Drug Administration research associateship.

‖ To whom correspondence should be addressed: Division of Cytokine Biology, Food and Drug Administration, HFM 505, 8800 Rockville Pike, Bethesda, MD 20892. Tel.: 301-496-0864; Fax: 301-402-1659.

[1] The abbreviations used are: Jak, Janus kinase; Stat, signal transducer and activator of transcription; SH2, *src* homology domain 2; GRR, γ response region; GAS, γ-activated site; SIE, serum-inducible element; ISRE, interferon-stimulated response element; DER, *Drosophila* epidermal growth factor receptor; PAGE, polyacrylamide gel electrophoresis; VBC, vanadate/peroxide-induced binding complex.

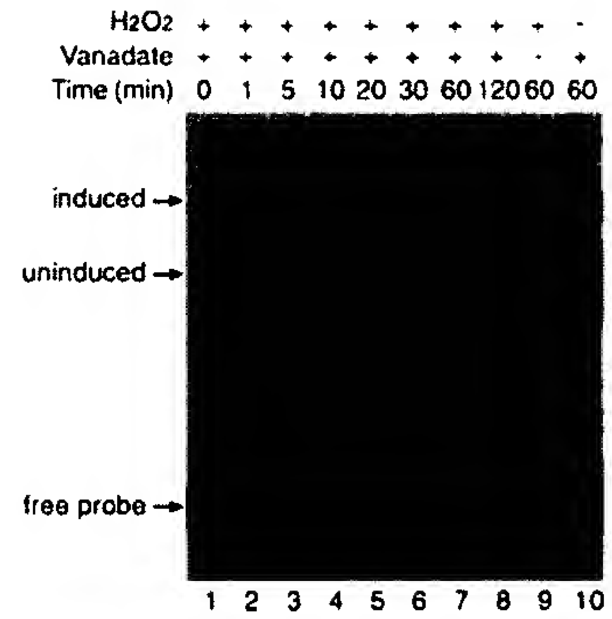


FIG. 1. **Vanadate/peroxide-inducible GRR binding complex in *Drosophila* Schneider 2 cells.** Whole cell lysates were made from untreated Schneider 2 cells (*lane 1*), cells treated with vanadate/peroxide for increasing amounts of time (*lanes 2–8*), or cells treated for 1 h with either hydrogen peroxide alone (*lane 9*) or vanadate alone (*lane 10*). Whole cell lysates from these cells were analyzed by electrophoretic mobility shift assay (as described under "Materials and Methods").

temperature. Where indicated, binding reactions were preincubated for 30 min with 5 mM phenyl phosphate, 5 mM sodium phosphate, or *Yersinia* protein tyrosine phosphatase in the presence or absence of 1 mM sodium orthovanadate or 3 mM sodium tungstate. Binding was analyzed on a 6% non-denaturing gel containing 2.5% glycerol in 0.25 × TBE (22.5 mM Tris, 22.2 mM boric acid, 0.5 mM EDTA). Gels were dried and subjected to autoradiography.

Protein Purification and SDS-PAGE Analysis—Soluble cell lysate from activated or control Schneider 2 cells was allowed to bind to heparin-agarose (Sigma) for 1 h at 4 °C. The agarose was washed with cell lysis buffer, and the vanadate/peroxide-induced binding complex (VBC) was eluted with a salt gradient from 0 to 300 mM NaCl in cell lysis buffer. Biotinylated GRR oligonucleotide was bound to streptavidin-agarose according to standard protocols. Partially purified VBC was allowed to incubate with GRR-agarose in the presence of 200 μg/ml salmon sperm DNA (Digene Diagnostics, Inc.) for 1 h at 4 °C. Affinity beads were washed 3 times with cell lysis buffer. Protein was solubilized using SDS-PAGE sample buffer. After denaturation, proteins were separated on a 4–20% acrylamide gel and transferred to polyvinylidene difluoride membranes (Immobilon, Millipore) by electroblotting. Blots were subjected to standard Western blotting procedures using biotinylated 4G10 anti-phosphotyrosine, streptavidin-conjugated horseradish peroxidase, and ECL (Amersham Corp.) for detection.

Southern Analysis—10 μg of normal human genomic DNA, 10 μg of *Drosophila* genomic DNA, or 2.5 μg of *Drosophila* genomic DNA was digested using *Eco*RI. Southern blot analysis, at a stringency reduced by 14 °C, was performed as described (11). The probe was generated by the polymerase chain reaction using oligonucleotide primers (5′-TACTGTGTTCATCATACTGTC-3′ and 5′-TGGAATGATGGATGCATCATGGCTT-3′), spanning nucleotides 1913–2446 of human Stat1α, and labeled by nick translation.

RESULTS AND DISCUSSION

Vanadate/Peroxide Treatment of Drosophila Schneider 2 Cells Induces Binding of a Specific Complex to the GRR of the Fc Receptor Promoter—We have previously shown that vanadate/peroxide will mimic the action of interferon γ on monocytes (5). We have also detected vanadate/peroxide stimulation of GRR-binding proteins in a number of different cell types, including HeLa, U266, THP1, U937, Daudi, and fibroblast lines.[2] Because *Drosophila* interferons have yet to be identified, we activated potential Stat-like proteins with vanadate/peroxide, thereby bypassing a ligand/receptor interaction. Treatment of Schneider 2 cells with vanadate/peroxide resulted in the formation of a complex that specifically bound to the GRR in a time-dependent manner (Fig. 1). Cells required 20 min of exposure to vanadate/peroxide for assembly of the complex as measured by electrophoretic mobility shift assays. DNA binding activity increased for up to 120 min of incubation (Fig. 1, *lanes 1–8*). Longer incubation times did not increase the intensity of the shift (data not shown). The intensity of the *Drosophila* shift was similar to the intensity of shifts detected in mammalian cells; however, the *Drosophila* shift complex migrated slightly faster on the native gel than did the mammalian shift complex (data not shown). The binding activity was induced in the presence of cycloheximide suggesting that new protein synthesis was not necessary for activity (data not shown). These findings suggest that the appearance of DNA binding activity is an early step in the activation of this pathway.

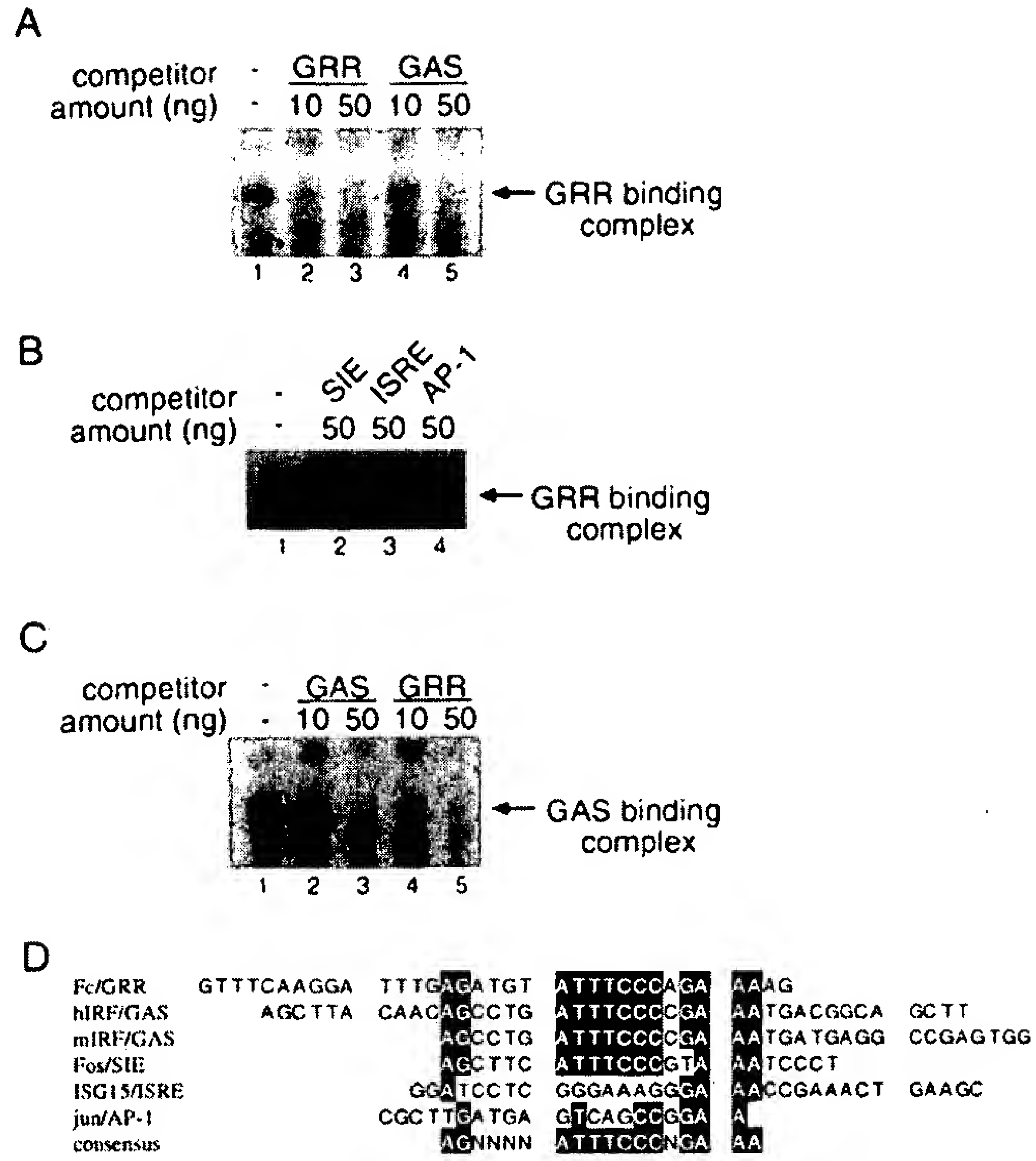


FIG. 2. **Vanadate/peroxide-inducible DNA binding complex binds specifically to GRR and GAS elements.** *A*, vanadate/peroxide-induced GRR binding complexes formed in the presence of a 10-fold (*lane 2*) or 50-fold (*lane 3*) excess of unlabeled GRR and a 10-fold (*lane 4*) or 50-fold (*lane 5*) excess of unlabeled GAS. *B*, vanadate/peroxide-induced GRR binding complexes formed in the presence of a 50-fold excess of unlabeled SIE (*lane 2*), a 50-fold excess of unlabeled ISRE (*lane 3*), or a 50-fold excess of unlabeled AP-1 (*lane 4*). *C*, vanadate/peroxide-induced GAS binding complexes formed in the presence of a 10-fold (*lane 2*) or 50-fold (*lane 3*) excess of unlabeled GAS and a 10-fold (*lane 4*) or 50-fold (*lane 5*) excess of unlabeled GRR. *D*, comparison of the enhancer sequences used for electrophoretic mobility shift assays and competitions. A consensus sequence was determined by comparing sequences that competed for the GRR binding complex.

To determine whether the induced complex bound specifically to the GRR, we performed competition experiments using specific unlabeled oligonucleotides. A 10-fold excess of either unlabeled GRR or a closely related γ-activated sequence (GAS) from the IRF1 promoter competed for labeled GRR binding (Fig. 2*A*, *lanes 2* and *4*), while only a slight decrease in binding was observed when a 50-fold excess of a high affinity SIE (13) was added (Fig. 2*B*, *lane 2*). Addition of either unlabeled ISRE or AP-1 did not effect the GRR binding complex (Fig. 2*B*, *lanes 3* and *4*). A DNA binding complex was also detected when a labeled IRF1/GAS element was used in the electrophoretic mobility shift assay. Addition of either unlabeled GRR or GAS to this binding reaction competed for the labeled GAS binding complex (Fig. 2*C*, *lanes 2–4*). These data suggest that the VBC has approximately the same affinity for the GAS and GRR sites

[2] A. C. Larner and D. S. Finbloom, unpublished data.

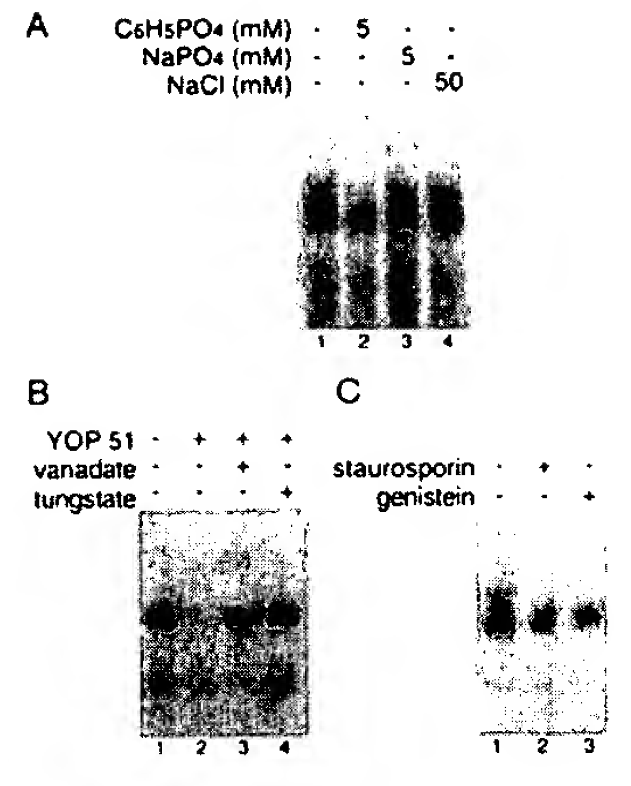


FIG. 3. **Tyrosine phosphorylation is necessary for maintenance of the shift complex.** *A*, whole cell lysates from vanadate/peroxide-stimulated Schneider 2 cells were untreated (*lane 1*), treated with 5 mM phenyl phosphate (*lane 2*), treated with 5 mM sodium phosphate (*lane 3*), or treated with 50 mM sodium chloride (*lane 4*). *B*, whole cell lysates from vanadate/peroxide-stimulated Schneider 2 cells were untreated (*lane 1*), treated with YOP51 (*lane 2*), or treated with YOP51 in the presence of 1 mM sodium orthovanadate (*lane 3*) or 3 mM sodium tungstate (*lane 4*). *C*, Schneider 2 cells were pretreated with no addition (*lane 1*), 500 nM staurosporin for 30 min (*lane 2*), or 30 μg/ml genistein for 30 min (*lane 3*). Cells were then treated with vanadate/peroxide for 60 min. Whole cell lysates from all experiments were analyzed by electrophoretic mobility shift assay.

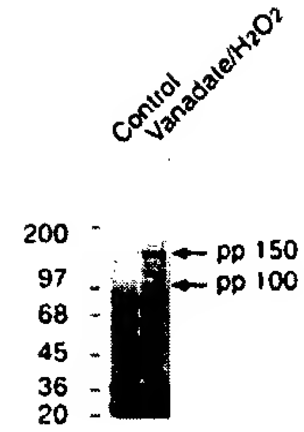


FIG. 4. **Vanadate/peroxide-stimulated tyrosine phosphorylation of two GRR-binding proteins.** Partially purified vanadate/peroxide-stimulated binding complex was bound to a GRR affinity column in the presence of salmon sperm DNA to compete for nonspecific DNA-binding proteins. Proteins adsorbed to the column were eluted in SDS sample buffer and separated by SDS-PAGE. The proteins were transferred to Immobilon and probed with 4G10. The two dominant proteins that were inducibly phosphorylated are indicated.

as judged by competition analysis but has a lesser affinity for the SIE sequence. Electrophoretic mobility shift assays supported these findings where the GAS and GRR elements showed strong binding and the SIE showed slight but detectable binding (data not shown).

The DNA elements that were bound by the VBC share a consensus ATTTCCCNGAAA core region that contains the GAS element described in the mammalian system (Fig. 2*D*) (2). The ISRE and AP-1 elements, which did not compete, lack this consensus sequence.

Tyrosine Phosphorylation Is Necessary for the Formation of the Binding Complex—It has been clearly established that the Stat-like proteins are phosphorylated on tyrosine residues in response to cytokine stimulation of cells and that this phosphorylation is necessary for an active complex to be assembled (1, 14–20). A functional SH2 domain of Stat1α and phosphorylation of tyrosine 701 are necessary for dimer formation in response to interferon γ (4). Phosphopeptides corresponding to sequences surrounding tyrosine 701 can block dimer formation. It has been speculated that this disruption is due to the SH2 domain of Stat1α binding to the phosphopeptide instead of binding to phosphotyrosine 701. Phenyl phosphate can also disrupt Stat complexes, presumably by competing for SH2 domain binding (20). Phenyl phosphate added to vanadate/peroxide-stimulated extracts of Schneider 2 cells inhibited the DNA binding activity of the VBC (Fig. 3*A*, *lane 2*), while equimolar amounts of sodium phosphate or a 10-fold higher concentration of sodium chloride had no effect on the DNA binding activity (Fig. 3*A*, *lanes 3* and *4*). The inhibition of DNA binding activity by phenyl phosphate can be reversed by removing the salt by dialysis (data not shown). These results are consistent with the *Drosophila* vanadate/peroxide-inducible complex containing a protein with an SH2 domain.

Other assays have been used in the mammalian system to confirm that Stat-like proteins are tyrosine-phosphorylated and that this phosphorylation is important for maintaining a DNA binding activity (1, 20, 21). These assays have included the treatment of activated extracts with the tyrosine-specific phosphatase, YOP51, and pretreatment of cells with the tyrosine kinase inhibitor, staurosporin (21). Treatment of a Schneider 2 cell lysate from vanadate/peroxide-treated cells with the tyrosine-specific phosphatase, YOP51, disrupted the DNA binding activity (Fig. 3*B*, *lane 2*). This inhibition was reversed by the addition of either 1 mM sodium vanadate or 3 mM sodium tungstate, which are inhibitors of tyrosine-specific phosphatases. The VBC was also inhibited by pretreatment of the Schneider 2 cells with the tyrosine kinase inhibitors, genistein or staurosporin (Fig. 3*C*, *lanes 2* and *3*). These data suggest that a tyrosine kinase is involved in the formation of the *Drosophila* DNA binding complex and that tyrosine phosphorylation is necessary for DNA binding activity.

Vanadate/Peroxide Induces the Tyrosine Phosphorylation of Two Major GRR-binding Proteins, pp100 and pp150—VBC was partially purified as described under "Materials and Methods" and adsorbed to a GRR oligonucleotide affinity resin. The loss of the VBC activity from the extract was monitored by electrophoretic mobility shift analysis. The DNA binding complex, which specifically adsorbed to the GRR column, was analyzed by Western blotting with 4G10, a monoclonal antibody that recognizes phosphotyrosine. Two inducibly phosphorylated proteins of approximate molecular mass of 100 and 150 kDa were detected (Fig. 4). Either of these sizes would be appropriate for a *Drosophila* Stat since the mammalian Stat proteins range in molecular mass from 84 kDa (Stat 4) to 113 kDa (Stat 2) (22, 23). Two minor bands of 120 and 130 kDa were not consistently detected and could be due to degradation of pp150.

Southern Analysis Demonstrates That a Drosophila Gene Exists Which Hybridizes to the Human Stat1α SH2 Domain Sequence—Based upon functional evidence presented here, we sought to determine whether the *Drosophila* genome harbors DNA sequences related to human Stat1α. Ten μg of normal human and *Drosophila* genomic DNA were digested with *Eco*RI, and the products were separated on agarose gels. To correct for differences in genomic complexity of both species, 2.5 μg of *Drosophila* genomic DNA were also analyzed. Southern blot analysis was performed using the SH2 domain coding sequence of human Stat1α as a probe. As shown in Fig. 5, three major bands of 3, 4.4, and 10 kilobases, respectively, were observed in *Drosophila* DNA in addition to several minor bands. The strongly hybridizing fragments were also detectable in 2.5 μg of *Drosophila* genomic DNA ensuring specificity of hybridization. These findings suggest that the *Drosophila* genome contains at least one and perhaps three genes related to the mammalian Stat1α SH2 domain.

In summary, we have identified a unique DNA binding activity in *Drosophila melanogaster*. This activity resembles the Stat-like activity that has been extensively characterized in the mammalian system. In both the mammalian and the *Drosoph-*

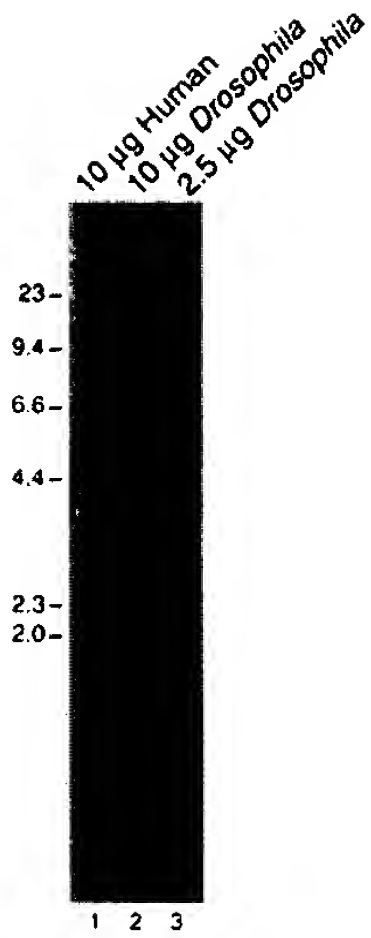


FIG. 5. **Stat1-related sequences identified in *Drosophila* genomic DNA by reduced stringency hybridization.** 10 µg of human genomic DNA (*lane 1*), 10 µg of *Drosophila* genomic DNA (*lane 2*), or 2.5 µg of *Drosophila* genomic DNA (*lane 3*) was digested with *Eco*RI and separated on a 0.8% agarose gel. The SH2 domain of human Stat1α was labeled by nick translation and used as a probe. The mobility of molecular weight standards (λ *Hin*dIII) is indicated on the *left*.

ila systems, vanadate/hydrogen peroxide treatment of cultured cells induces a specific GRR binding complex whose formation is dependent upon tyrosine phosphorylation. Detection of Stat1α-related sequences in the *Drosophila* genome raises the possibility that, as in the mammalian system, Stat1α-like proteins are responsible for this activity. Work is in progress to obtain the protein sequence and to clone the cDNA encoding such novel Stat proteins.

Work presented here is important to the understanding of the mammalian Jak/Stat signaling pathways. *Drosophila* can provide a genetic model of Jak/Stat activation and give insight into the significance of conserved protein sequences. This work is also important for *Drosophila* development because, like the *Drosophila* Jak protein, the Stat-like protein may play a role in signal transduction during stripe formation.

REFERENCES

1. Larner, A. C., David, M., Feldman, G. M., Igarashi, K., Hackett, R. H., Webb, D. A. S., Sweitzer, S. M., Petricoin, E. F., III, and Finbloom, D. S. (1993) *Science* **261**, 1730–1733
2. Darnell, J. E., Jr., Kerr, I. M., and Stark, G. R. (1994) *Science* **264**, 1415–1421
3. Shuai, K., Schindler, C., Prezioso, V. R., and Darnell, J. E., Jr. (1992) *Science* **258**, 1808–1812
4. Shuai, K., Horvath, C. M., Tsai Huang, L. H., Qureshi, S. A., Cowburn, D., and Darnell, J. E., Jr. (1994) *Cell* **76**, 821–828
5. Igarashi, K., David, M., Larner, A. C., and Finbloom, D. S. (1993) *Mol. Cell. Biol.* **13**, 3984–3989
6. Binari, R., and Perrimon, N. (1994) *Genes & Dev.* **8**, 300–312
7. Duffy, J. B., and Perrimon, N. (1994) *Dev. Biol.* **166**, 380–395
8. Perrimon, N. (1993) *Cell* **74**, 219–222
9. Raz, E., Eyal, S. D., and Shilo, B. Z. (1991) *Genetics* **129**, 191–201
10. Shilo, B. Z. (1992) *FASEB J.* **6**, 2915–2922
11. Kraus, M. H., and Aaronson, S. A. (1991) *Methods Enzymol.* **200**, 546–556
12. Deleted in proof
13. Wagner, B. J., Hayes, T. E., Hoban, C. J., and Cochran, B. H. (1990) *EMBO J.* **9**, 4477–4484
14. Gutch, M. G., Daly, C., and Reich, N. C. (1992) *Proc. Natl. Acad. Sci. U. S. A.* **89**, 11411–11415
15. Schindler, C., Shuai, K., Prezioso, V. R., and Darnell, J. E., Jr. (1992) *Science* **257**, 809–813
16. Kotanides, H., and Reich, N. C. (1993) *Science* **262**, 1265–1267
17. Shuai, K., Ziemiecki, A., Wilks, A. F., Harpur, A. G., Sadowski, H. B., Gilman, M. Z., and Darnell, J. E., Jr. (1993) *Nature* **366**, 580–583
18. Rothman, P., Kreider, B., Azam, M., Levy, D., Wegenka, U., Eilers, A., Decker, T., Horn, F., Kashleva, H., Ihle, J., and Schindler, C. (1994) *Immunity* **1**, 457–468
19. Zhong, Z., Wen, Z., and Darnell, J. E., Jr. (1994) *Science* **264**, 95–98
20. David, M., Romero, G., Zhang, Z., Dixon, J. E., and Larner, A. C. (1993) *J. Biol. Chem.* **268**, 6593–6599
21. Fu, X. Y. (1992) *Cell* **70**, 323–335
22. Yamamoto, K., Quelle, F. W., Thierfelder, W. E., Kreider, B. L., Gilbert, D. J., Jenkins, N. A., Copeland, N. G., Silvennoinen, O., and Ihle, J. N. (1994) *Mol. Cell. Biol.* **14**, 4342–4349
23. Fu, X-Y., Kessler, D. S., Veals, S. A., Levy, D. E., and Darnell, J. E., Jr. (1990) *Proc. Natl. Acad. Sci. U. S. A.* **87**, 8555–8559

Methods in Enzymology

Volume 200

Protein Phosphorylation

Part A

Protein Kinases: Assays, Purification, Antibodies, Functional Analysis, Cloning, and Expression

EDITED BY

Tony Hunter
Bartholomew M. Sefton

MOLECULAR BIOLOGY AND VIROLOGY LABORATORY
THE SALK INSTITUTE
SAN DIEGO, CALIFORNIA

ACADEMIC PRESS, INC.
Harcourt Brace Jovanovich, Publishers
San Diego New York Boston
London Sydney Tokyo Toronto

Plus/Minus Screening

The screening of PCR libraries generated with the oligonucleotides *PTK*1 and *PTK*2 has been based on a random accumulation of sequence data, followed by subsequent selection of sequences of interest. While in the early phase of this work a high proportion of clones contained new *PTK*-related sequences, later experiments proved to be somewhat less fruitful. Moreover, the selection of clones for sequencing was a purely random process and was uninformative with regard to other potentially important selection criteria, such as expression pattern. An improvement in this respect has been the opportunity to screen PCR libraries for *PTK*-related sequences which show a tissue-specific or developmentally regulated pattern of expression. The PCR library generated from the mRNA source of interest is screened, in duplicate, with ^{32}P-labeled PCR product from which the library was constructed. The filters are then stripped and reprobed with a PCR probe generated in an identical fashion from an mRNA source other than that used to construct the PCR library. In this way we have been able to generate differentially expressed *PTK*-related clones expressed in epithelial cells but not fibroblasts (A. Ziemiecki and A. F. W., unpublished data). Although this procedure has been employed successfully we have not tested the limits of its sensitivity, nor its utility in other situations, such as, for example, a differentiation system.

[46] Detection and Isolation of Novel Protein-Tyrosine Kinase Genes Employing Reduced Stringency Hybridization

By M. H. KRAUS and S. A. AARONSON

Introduction

Protein-tyrosine kinases (PTKs) are encoded by a conserved family of ancestrally related genes that have segregated during evolution within the larger family of protein kinases.[1] They share a catalytic domain encoding 250–300 amino acid residues that represents the region of most extensive structural conservation and harbors the intrinsic enzymatic activity. Individual members in distinct subfamilies of both receptor-like and cytoplasmic tyrosine kinases share closer structural and functional homology with each other than with other protein kinases. The higher degree of structural homology among functionally more closely related PTKs pro-

[1] S. K. Hanks, A. M. Quinn, and T. Hunter, *Science* **241**, 42 (1988).

vides a basis for the identification of novel members with predictable functional properties due to their association with individual PTK subfamilies. Based upon concurrent nucleotide sequence conservation, the search for novel *PTK* genes can be directed toward a specific subfamily without detection of more distantly related *PTK* genes, due to cross-hybridization of genomic exon sequences of novel members under quantitatively reduced hybridization stringencies with *PTK* domain probes of known subfamily members. Analyzing normal genomic DNA representing *PTK* genes at single-copy level permitted detection of exon sequences encoding closely related novel PTKs independent of their expression pattern.

Novel genomic restriction fragments are cloned and the most conserved region is mapped by Southern blot analysis. Thus, exons with highest homology to the probe are identified and their structure is determined by nucleotide sequence analysis, establishing the degree of structural relatedness and exon organization of a novel related tyrosine kinase gene. Furthermore, exon-containing probes can be derived for expression analysis as well as the isolation of cDNA clones of novel *PTK*s. Employing probes from the TK domain of v-*erbB*, v-*fms*, and v-*abl* we have isolated with *erbB-2*[2] and *erbB-3*[3] two novel receptor-like tyrosine kinases in the *erbB*/EGF-R family, a second PDGF-R[4] in the *fms*/CSF1-R family, as well as a gene closely related to *abl*,[5] respectively in the human genome.

Genomic Southern Blot Analysis

The reliable detection of novel *PTK* genes at reduced stringency requires a high sensitivity and specificity of genomic Southern blot analysis.[6] Using a slightly modified procedure we detect human single-copy genomic DNA fragments with a completely matching probe at high stringency after 2–4 hr of film exposure at $-70°$ in the presence of intensifier screens.

1. Ten micrograms genomic DNA is restricted in a final volume of 200–400 μl. To monitor restriction, bacteriophage λ DNA is incubated with a small aliquot of the main reaction at identical DNA/enzyme unit ratio.
2. After complete restriction, DNA samples are purified with an equal

[2] C. R. King, M. H. Kraus, and S. A. Aaronson, *Science* **229**, 974 (1985).
[3] M. H. Kraus, W. Issing, T. Miki, N. C. Popescu, and S. A. Aaronson, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 9193 (1989).
[4] T. Matsui, M. Heidaran, T. Miki, N. Popescu, W. LaRochelle, M. Kraus, J. Pierce, and S. Aaronson, *Science* **243**, 800 (1989).
[5] G. D. Kruh, C. R. King, M. H. Kraus, N. C. Popescu, S. C. Amsbaugh, W. O. McBride, and S. A. Aaronson, *Science* **234**, 1545 (1986).
[6] E. M. Southern, *J. Mol. Biol.* **98**, 503 (1975).

volume of buffered phenol/CIA (chloroform/isoamyl alcohol, 24 : 1) and precipitated from 1 *M* ammonium acetate with 2.5 vol cold ethanol on dry ice for 20 min. The pelleted DNA is carefully washed with cold 75% (v/v) ethanol, lyophilized, and following solubilization in 30 μl 1 × sample buffer loaded on a 0.8–1.2% (w/v) horizontal agarose gel (20 × 25 cm; i.e., BRL, Gaithersburg, MD). Electrophoresis proceeds overnight at 35 V in a Tris-acetate gel/buffer system (1 × E) containing 0.5 μg/ml ethidium bromide with buffer recirculation.

40 × E buffer: 1.6 *M* Tris, 0.8 *M* sodium acetate, 40 m*M* EDTA; adjust pH to 7.2 with acetic acid

10 × sample buffer: 0.2 *M* Tris-acetate, pH 7.5, 0.02 *M* EDTA, 1% (w/v) SDS, 50% (v/v) glycerol, 0.3% (w/v) Bromphenol Blue

3. Following gel photography, the gel is irradiated for 2 min with 302-nm UV light to improve transfer of larger DNA fragments, and trimmed to 20 × 20 cm. The DNA is denatured by two consecutive gel treatments of 15 min in 0.5 *M* NaOH/1.5 *M* NaCl and equilibrated with two treatments of 15 min each in 1 *M* ammonium acetate/0.02 *M* NaOH. Capillary transfer in this solution to standard 0.45-μm nitrocellulose proceeds overnight with the bottom of the gel facing the membrane. After transfer, the filter is baked for 2 hr at 80° under vacuum.

4. Standard high-stringency hybridization is conducted in 5 × SSC (1 × SSC = 0.15 *M* NaCl, 0.015 *M* sodium citrate, pH 7) and 50% formamide at 42°, which establishes conditions 20–25° below the T_m (melting point) of a completely matched DNA hybrid. For hybridization the membrane is placed in a sealing bag, wetted in two-thirds of the final volume (0.075 ml/cm^2 filter area), and probe is added in the remaining third of hybridization solution. The bag is sealed and after thorough mixing of the solutions placed between two glass plates in a 42° water bath for 8–16 hr.

Hybridization solution (1 ×): 5 × SSC, 10% (w/v) dextran sulfate, 2.5 × Denhardt's solution, 10 m*M* Tris pH 7.4, 50 μg/ml sheared and boiled salmon sperm DNA, 50% (v/v) formamide, 2–5 × 10^6 cpm/ml at a DNA concentration of <5 ng/ml

Hybridization buffer containing SSC, dextran sulfate, Denhardt's, and Tris is prepared as 2 × stock solution by dissolving dextran sulfate powder in SSC while stirring prior to adding the other components. Nonradioactive and radioactive hybridization solutions are prepared by adding salmon sperm DNA to the formamide, and purified probe is added to the latter solution. Both are then incubated 5 min at 50°, thoroughly mixed, and hybridization buffer is added to 1 × final concentration. This step will ensure sufficient denaturation of both carrier and labeled probe DNA.

For reduced stringency hybridization, the volume is adjusted with water following denaturation in formamide. High specific activity probes of $>10^9$ cpm/μg DNA can be routinely obtained with commercially available nick translation (i.e., Amersham, Arlington Heights, IL) or random primer kits. Under these conditions "prehybridization" of the filter is not required, as long as the filter is prewetted in nonradioactive hybridization solution prior to adding the probe.

5. After hybridization, filters are washed three times for 5 min each in 2× SSC, 0.01% SDS at room temperature. Two high-stringency washes of 20 min each are then conducted at 50° in 0.1× SSC/0.1% SDS and 0.1× SSC, respectively.

Detection of Novel Tyrosine Kinase Gene Fragments by Genomic Hybridization Analysis at Reduced Stringencies

Cross-hybridization of related tyrosine kinase gene fragments is accomplished by reduced stringency hybridization facilitating hybrid formation and stability between partially mismatched sequences. While theoretical aspects of the formation of DNA hybrids and sequence homology have been extensively discussed in the literature,[7] most of these findings derive from liquid hybridization and may vary to some extent in filter hybridization where the template DNA is immobilized.

Thus, to determine experimentally the optimal stringency for cross-hybridization of a novel related tyrosine kinase gene, it proves useful to subject replica nitrocellulose filters containing genomic DNA to hybridization under stringencies incrementally decreased by 7°. Discrete intermediate levels of reduced stringency (Table I) may be useful in some cases. Since stringency is presumed to affect predominantly the formation of hybrids during hybridization and the stability of hybrids during washing, both hybridization and washing are conducted under equally reduced stringencies. Experimentally these are controlled by the formamide concentration during hybridization and salt concentration during washing, respectively, with otherwise identical conditions (Table I).

It is helpful to relate stringency in temperature degrees (°C) assuming that decrease of formamide concentration by 1% increases the T_m of the hybrid by 0.7° and a logfold increase in ionic strength raises the stability of a hybrid by 18.5°.[7,8] Thus, considering incubation temperature (T_i),

[7] M. L. M. Anderson and B. D. Young, *in* "Nucleic Acid Hybridisation" (B. D. Hames and S. J. Higgins, eds.), p. 73. IRL Press, Oxford, 1985.

[8] G. A. Beltz, K. A. Jacobs, T. H. Eickbush, P. T. Cherbaş, and F. C. Kafatos, *in* "Methods in Enzymology" (R. Wu, L. Grossman, and U. Moldave, eds.), p. 266. Academic Press, New York, 1983.

TABLE I
STRINGENCY REDUCTION[a]

Stringency reduction (ΔT_s) (°C)	Hydridization (42°/5× SSC)	Washing (50°)
0	**50% FA**	**0.1× SSC → 0.02 *M* Na⁺**
−3.5	45% FA	0.15× SSC → 0.03 *M* Na⁺
− 7	40% FA	0.25× SSC → 0.04 *M* Na⁺
−10.5	35% FA	0.4× SSC → 0.08 *M* Na⁺
−14	30% FA	0.6× SSC → 0.12 *M* Na⁺
−17.5	25% FA	1× SSC → 0.2 *M* Na⁺
−21	20% FA	1.5× SSC → 0.3 *M* Na⁺

[a] ΔT_s indicates reduction of stringency relative to high-stringency conditions (bold type). FA, Formamide; 1× SSC = 0.15 *M* NaCl, 0.015 *M* sodium citrate, pH 7. For each stringency, hybridization is conducted at 42° in 5× SSC, while washing occurs at 50°.

formamide concentration (FA), and ionic strength (μ), experimental stringency conditions relative to high-stringency hybridization can be approximated as

$$T_s = T_i + 0.7(\%\ \mathrm{FA}) - 18.5 \log(\mu/\mu_0)$$

where T_s is the stringency expressed as temperature degrees, T_i the incubation temperature, and μ the ionic strength, while μ_0 represents the ionic strength during hybridization conditions equalling 1 *M* Na^+ (5× SSC). At these high salt concentrations the hybrid stability is maximal and relatively unaffected by variations of ionic strength (log 1 = 0). Under high-stringency conditions the estimated T_s of the hybridization ($T_s = 42° + 0.7° \times 50 - 18.5 \times \log 1$) is 77° and the T_s during washing ($T_s = 50° + 0.7° \times 0 - 18.5° \times \log 0.02$) is 81°. Reducing the formamide concentration by 10% results in an estimated stringency of $T_s = 70°$, yielding a stringency reduction of $\Delta T_s = -7°$ in the hybridization. Equivalent reduction of the washing stringency requires an increase of the salt concentration to 0.25 × SSC (0.05 *M* Na^+). Table I lists experimental conditions with the respective estimated reduction in stringency achieved simultaneously during hybridization and washing. Since the rate of filter hybridization is not affected by formamide concentrations in the range from 50 to 30% and only slightly reduced at 20% formamide,[7] hybrid formation at different stringencies mainly depends on the degree of mismatches between probe and template DNA, eliminating hybridization kinetics as a major variable. Under these experimental conditions, we observe comparable signal/noise ratios at both high and reduced stringency.

The extent of stringency reduction required to detect a related se-

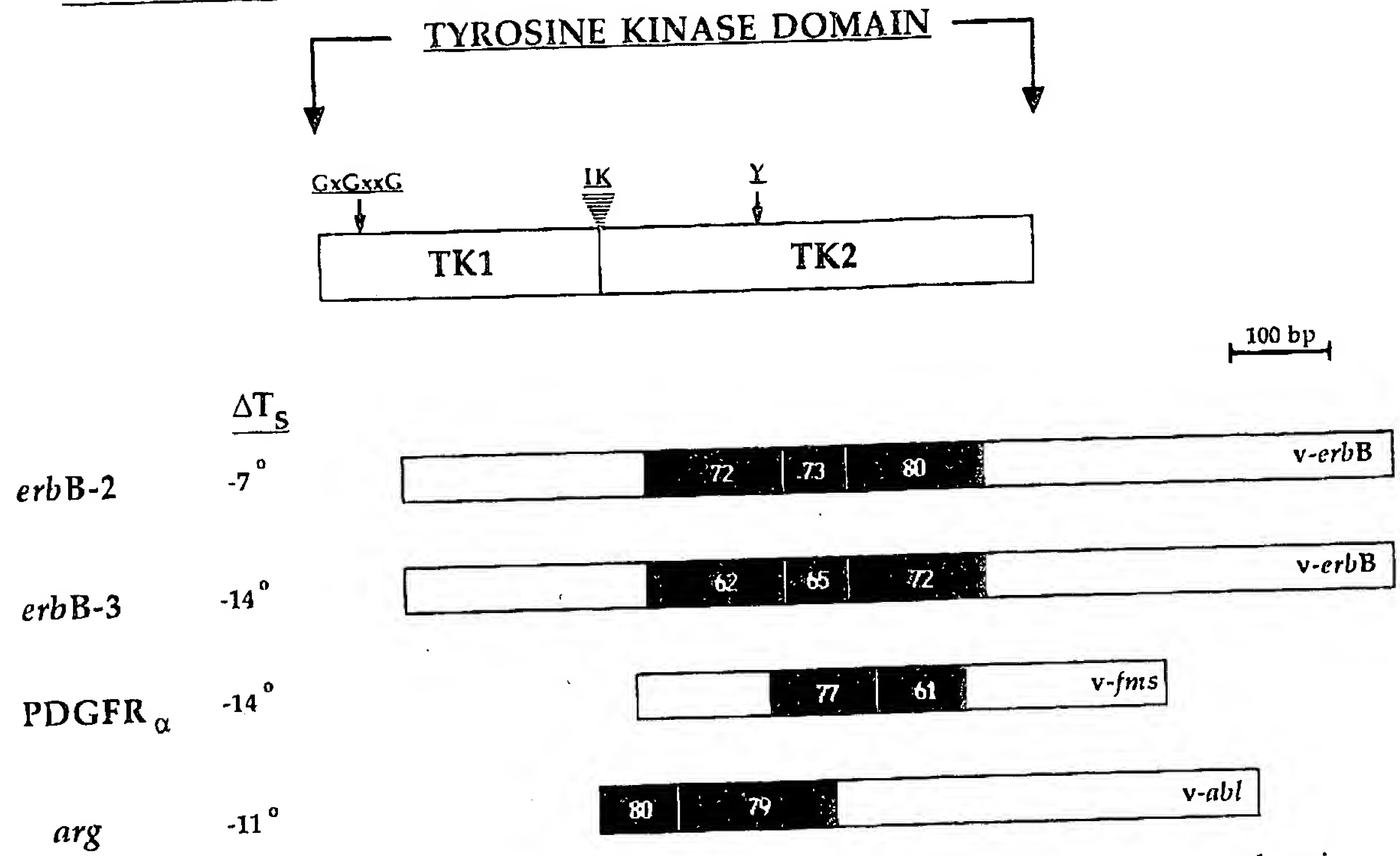


FIG. 1. Structural relationship of exons encoding novel receptor-like and cytoplasmic tyrosine kinases detected by reduced stringency hybridization. The junction between tyrosine kinase subdomains TK1 and TK2 is defined by a stretch of less well-conserved coding sequence, the interkinase region (IK), present in the *fms*/PDGFR family. The positions of ATP-binding site (GxGxxG), interkinase region (IK), and a tyrosine residue homologous to the autophosphorylation Tyr-416 (Y) of v-*src* are indicated. ΔT_s represents the reduction of stringency required for detection of related tyrosine kinases. Percentage nucleotide sequence identities of individual exons (solid boxes) with the homologous regions of the utilized probes (open bars) are indicated.

quence can be correlated to some degree with the nucleotide sequence identity score. Hybridization efficiency and hybrid stability of related sequences also depend on the distribution of nucleotide sequence homology and contiguous stretches of sequence identity with the probe (Fig. 1). Furthermore, in using genomic DNA, experimental conditions are influenced by the intron–exon structure. A closer homology between related tyrosine kinases tends to coincide with conserved exon structure. For instance, three exons that have been characterized in the tyrosine kinase domains of both *erbB-2* and *erbB-3* genes exhibit identical splice junctions, while these splice borders differ from those of tyrosine kinases of other subfamilies (Fig. 1). Thus, due to the conservation of nucleotide sequence and exon structure, the approach enhances the probability for detection of gene fragments most closely related to the gene from which the probe is derived. Figure 1 compares nucleotide sequence identity scores between exons of novel *PTK*s cross-hybridizing with the probe and the reduction of stringency required for the detection of single-copy gene

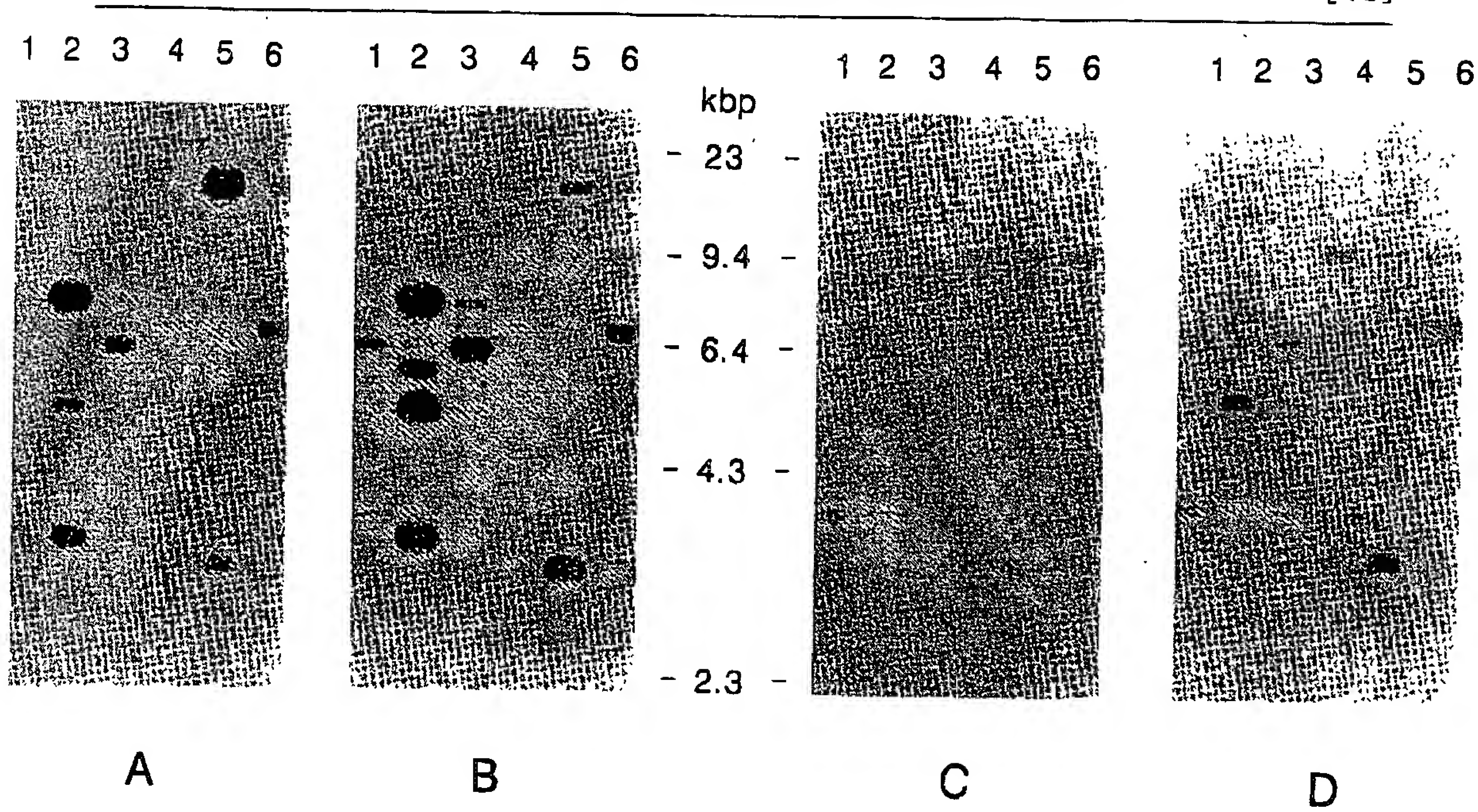


FIG. 2. *erbB*-Related sequences in the human genome. Replica Southern blots containing *Eco*RI (lanes 1–3)- or *Sac*I (lanes 4–6)-cleaved normal human genomic (1, 4), MDA-MB468 (2, 5), and SK-BR-3 (3, 6) DNAs were hybridized with v-*erbB* at intermediate ($\Delta T_s = -7°$) (A) and reduced ($\Delta T_s = -14°$) (B) or with an exon-specific *erbB*-3 probe at high (C) and reduced stringencies ($\Delta T_s = -14°$) (D). The migration positions of novel *erbB*-related *Eco*RI and *Sac*I restriction fragments identified under reduced stringency hybridization with v-*erbB* as a probe are indicated by arrowheads in (B).

fragments in normal DNA. This suggests that a reduction of stringency by 14° is adequate to detect related tyrosine kinase exons sharing in the range of 61 to 80% nucleotide sequence identity with the probe. In the EGF-R family, single gene copies of the more closely related *erbB-2* were detected at only 7° stringency reduction (Fig. 1).

A typical experimental approach is illustrated in Fig. 2. Replica nitrocellulose filters are generated containing several restriction digests of normal genomic DNA (lanes 1 and 4). The use of multiple enzyme digests increases the probability of detecting a novel *PTK* fragment distinct in size from those of the cognate gene. Furthermore, it is possible to exclude restriction fragment length polymorphisms. In this example, conducted for the identification of *erbB-3*,[3] we controlled for the gene fragments of known family members by their amplification in tumor cell DNAs. Alternatively, restriction fragments of known family members are identified by high-stringency hybridization with a perfectly matched probe. Using v-*erbB* as a probe at a stringency reduced by 7° (Fig. 2A), only restriction fragments specific to the EGF-R or *erbB-2* genes hybridized as determined by their amplification in MDA-MB468 (lanes 2 and 5) or SK-BR-3 (lanes 3 and 6), respectively. An additional 7° reduction of stringency

to $\Delta T_s = -14°$ facilitated the detection of a novel v-*erbB*-related gene distinct from EGF-R and *erbB-2* genes. It was most noticeable in the *SacI* digest as a 9-kbp restriction fragment (Fig. 2B, lanes 4–6), but was also discerned as a 3.7-kbp *Eco*RI fragment (Fig. 2B, lanes 1–3). The specific detection of a novel *erbB*-related gene was inferred by absence of hybridization under the higher stringency conditions and the lack of gene amplification in MDA-MB468 and SK-BR-3 cells.

Evaluation of Southern blots and subsequent genomic cloning are facilitated by choosing restriction enzymes and probes such that single or few bands for individual *PTK* genes are generated. If exon structure and nucleotide sequence conservation are known for a subfamily of interest, a shorter probe is preferred, as the probability of multiple hybridizing fragments decreases for each member. Otherwise, longer probes spanning several exons are preferable, since the region of relatively highest conservation within the TK domain varies among distinct families (Fig. 1). We have successfully utilized probes extending from a single exon (135 bp) to most of the tyrosine kinase domain, respectively. Endonucleases cleaving the coding sequence in the probe region should be avoided. This is underscored by comparison of the *SacI* with the EcoRI digest in Fig. 2 using a 1.1-kbp v-*erbB* probe at reduced stringency. The presence of the novel 3.7-kbp *Eco*RI *erbB-3* fragment is obscured by four EGF-R and one *erbB-2* fragments in the range of 3–8 kbp, whereas the 9-kbp *erbB-3*-specific *Sac*I fragment is readily distinguished from two EGF-R- and one *erbB-2*-specific fragments.

Characterization of Novel Tyrosine Kinase Exons

For further characterization, a novel related restriction fragment is cloned from a genomic library constructed by standard methodologies[9] in bacteriophage λ. At this step it is helpful to enrich the library source DNA for the novel restriction fragment (i.e., sucrose gradient centrifugation, DEAE membrane elution from agarose gel) with the attempt to exclude restriction fragment sizes of the cognate genes from the library. The clones of interest are identified by *in situ* hybridization at the reduced stringency initially determined by genomic Southern blot hybridization.

1. The bacteriophages are propagated according to standard procedures in 8 ml 0.7% top agarose in 150-mm plates on a 1.1% bottom agar layer at a density and plaque size that ensure that plaques are separated on the plates. After lifting the filters are air dried (15 min) and treated on 3MM Whatman sheets with the absorbed bacteriophages facing upward.

[9] T. Maniatis, E. F. Fritsch, and J. Sambrook, "Molecular Cloning: A Laboratory Manual." Cold Spring Harbor Lab., Cold Spring Harbor, New York, 1982.

Following denaturation (1 min) in 0.5 *M* NaOH/1.5 *M* NaCl the filters are briefly drained of excess alkali on dry 3MM paper, equilibrated with 1 *M* NH_4OAc/0.02 *M* NaOH (5 min), air dried, and baked for 2–3 hr at 80° in a vacuum oven.

2. For an efficient transfer of colonies of plasmid-based libraries from the plates onto nitrocellulose filters, 1.1% agar plates are used, where the melted agar has been cooled to 45° prior to pouring the plates. The marked filters are lifted gently, and with the colony side facing upward instantly placed onto 3MM paper saturated with denaturing solution. Sufficient lysis can be visually monitored, as the colonies become translucent (1–2 min). The filters are then treated essentially as described above. To minimize background, the filters following treatment with 1 *M* NH_4OAc/0.02 *M* NaOH are pressed while still wet between dry sheets of 3MM paper by firmly rolling a 25-ml pipette across the sandwich. By peeling the nitrocellulose filter off the 3MM paper, excess bacterial debris will adhere to the 3MM paper whereas DNA remains bound to the nitrocellulose membrane.

3. Hybridization and washing conditions are essentially as described for genomic Southern blot hybridization. Hybridization is conducted in 150-mm Petri dishes. Radioactive and nonradioactive hybridization solutions are prepared as described earlier and poured into separate dishes. Filters are individually wetted in cold hybridization solution and submerged in hybridization solution containing the probe ($1–3 \times 10^6$ cpm/ml). The covered dish enclosed in a bag or β radiation-safe container is rotated slowly (<100 rpm) in a 42° orbital shaker for overnight hybridization. As a guideline, 30 large round filters can be hybridized in 1 dish containing ~100 ml of radioactive hybridization solution. An additional 100 ml of nonradioactive hybridization solution is required to wet the nitrocellulose filters.

4. Washing of the filters is carried out in 4-liter covered glass beakers throughout. As many as 50 filters can be processed in 1 container. For even washing, the filters must remain in motion and submerged in the solution. A volume of 1–1.5 liters of washing solution for each beaker is adequate. Four room-temperature washes of 15–20 min each are followed by two final stringency washes of 30 min each. It is advisable to monitor the actual temperature in the solution during the final stringency washes at 50° in order to ensure correct stringencies.

5. Clones for the cognate genes can be distinguished by hybridizing replica filters from the same plates with probes specific for those genes under high strigency. More conveniently, 1 μl of first cycle positive phage stocks are applied to gridded plates by piercing the surface of the top agarose that contains susceptible bacterial cells with a pointed pipette tip. Phage lysis will occur after a 10- to 12-hr incubation at 37°. These plates

are then used for the generation of replica nitrocellulose filters. The latter approach for distinguishing clones of the cognate genes is particularly useful, when many first round positives require further analysis.

For the characterization of *erbB-3*, a normal human genomic DNA library enriched for 8- to 12-kbp *Sac*I restriction fragments was constructed in bacteriophage. Among 4×10^5 recombinants, 29 positives were identified using v-*erbB* as a probe under reduced hybridization stringency. One positive represented an EGF-R clone, whereas none was detected for *erbB-2*. Several positives plaques are purified and the phage DNAs are subjected to restriction and hybridization analysis at reduced stringency in order to map the region of homology within the phage DNA insert. Hybridizing insert fragments of suitable size (preferably 1 kpb or less) are then subcloned in plasmid vectors with high replication rate (pUC, Bluescript, etc.) and subjected to nucleotide sequence analysis by the dideoxy chain termination method using supercoiled DNA as template.

Gene-specific probes from the genomic clone can be used to investigate at high-stringency mRNA expression in various sources. In addition, genetic alterations including gene amplification or rearrangement as well as altered expression for a novel tyrosine kinase gene can be searched for in human disease. The predicted amino acid sequence of exons can be used to design peptides for the generation of polyclonal antisera in efforts to study the encoded gene product. For structural characterization and functional studies involving *in vitro* expresssion of the protein, the complete coding sequence can be isolated by cDNA cloning using gene-specific exon-containing probes at high stringency. The initial expression analysis is useful in detection of sources with relatively high transcript levels for cDNA cloning.

Utilizing a single genomic exon probe of *erbB-3* (Fig. 1, downstream exon) we rehybridized the initial genomic Southern blot under both high and reduced stringencies. This probe detected under high stringency the *erbB-3*-specific genomic fragments initially identified by v-*erbB* under reduced stringency (Fig. 2C). As expected at the same level of stringency reduction, the *erbB-3*-specific exon probe recognized only the homologous EGF-R- and *erbB-2*-specific gene fragments in addition to the endogenous restriction fragments (Fig. 2D), corroborating the specificity as well as reproducibility of this approach.

Discussion

In this chapter, we have summarized a general approach for detecting novel *PTK*-related genes in genomic DNA. Based upon knowledge of structural conservation in a distinct *PTK* subfamily, the approach can be

optimized for specific tasks by choice of restriction enzymes, probe design, and genomic DNA source. Several other approaches (see Chapters 44, 45, 47 in this volume), including reduced stringency hybridization of cDNA libraries with TK probes or degenerate oligonucleotides, have been employed in the successful isolation of novel *PTK*s. More recently, the polymerase chain reaction with degenerate oligonucleotides as primers has been utilized to isolate novel TK-coding sequences from DNA complementary to mRNA. Finally, screening of cDNA expression libraries with anti-phosphotyrosine antibodies has facilitated the successful isolation of *PTK*s based upon their expression at the protein level.

Approaches utilizing RNA as a source either combine criteria for structural relationship and levels of expression of a novel *PTK* or are strictly based on protein function. In the latter case this might allow the isolation of *PTK*s lacking sufficient sequence conservation for detection by the other methods described. Those techniques involving RNA as the source have the potential advantage of providing a more rapid determination of larger portions of coding sequence than the genomic method. On the other hand, the search for novel *PTK*s in mRNA-derived sources could be biased by the expression pattern. For example, a highly conserved, novel *PTK* could be missed, if not sufficiently expressed in a particular cell source, or inefficiently converted to cDNA due to its structure.

Since normal genomic DNA contains a complete set of genes encoding PTKs, the genomic approach is most comprehensive in the search for structurally conserved genes. Moreover, titration of stringency reduction on the genomic Southern blot can provide information about the number of different genes detected at a given stringency. A potential disadvantage of this method is the detection of nonfunctional genes, although in our search for novel *PTK*s we have not encountered such genes. While an intermediate genomic cloning step is required, knowledge of the expression pattern of a novel *PTK* prior to cDNA cloning may also be critical to the subsequent isolation of its complete coding sequence from the appropriate cDNA libraries. Hence, each of the described methods entails certain advantages, and the method of choice is determined by the specific goal. While those methods utilizing cDNAs may be particularly valuable in the isolation of tissue-specific expressed novel *PTK*s or molecules of more divergent genetic structure, genomic identification of novel *PTK*s appears most suitable for the systematic search for structurally related PTK coding sequences.

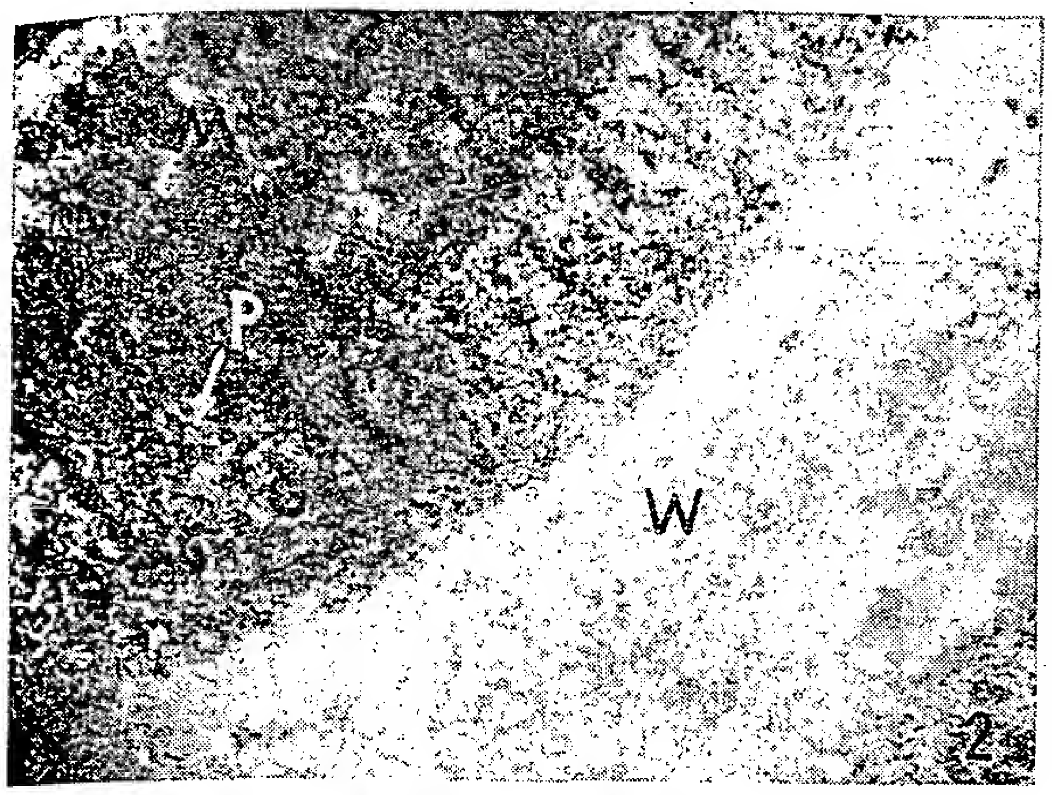


Fig. 2. Cerebellar cortex of the same rat as Fig. 1. Unstained autoradiogram of a cryostat section. M, Molecular layer; P, Purkinje cells (non-reacting); G, granular layer; W, white matter.

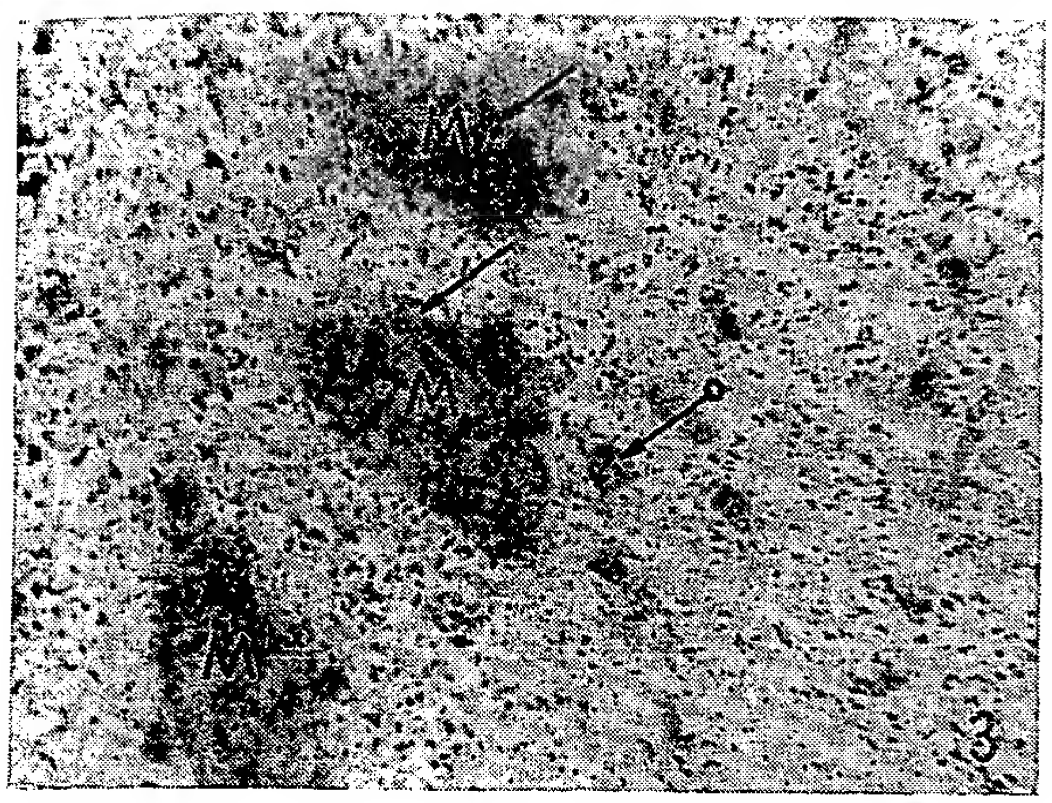


Fig. 3. Lumbar spinal cord (anterior horn) of a rat, 45 min after a single injection of TSC-^{14}C. Bouin-fixed and embedded specimen, counterstained with haematoxylin and eosin. Motoneurones (M) are non-reacting; capsular glial cells (arrow) and astrocytes in the grey matter (arrow with a circle) contain numerous grains.

binding by non-reacting structures might indicate those cellular elements that do not exert GAD activity. Such an approach is actually an adaptation of the principle described by Ostrovsky and Barnard[5].

Rats weighing 160 g were injected intraperitoneally with 5–20 mg/kg thiosemicarbazide-^{14}C (specific activity 10 mCi/mmole, obtained from the Central Isotope Institute, Budapest). Characteristic convulsions and lethal "jumps" appeared 45–120 min after injection. The animals were killed by decapitation and samples of the central nervous system (brain, medulla, cerebellum, spinal cord) as well as other tissues were either fixed on Bouin's solution and prepared for autoradiography in the usual way or were not fixed but frozen with dry ice, cut on a cryostat and applied to slides pre-coated with Kodak *AR*–10 stripping film, emulsion side up. The exposure time was 3–8 weeks; autoradiograms were developed in Kodak autoradiographic developer and some were counterstained with haematoxylin and eosin.

Fig. 1 shows a coronal section of the brain of a rat given two intraperitoneal injections of 10 mg/kg TSC-^{14}C, with an interval of 40 min between them, and then killed 90 min after the first injection. The heaviest reaction is confined to the hippocampus and fascia dentata chiefly in the layer of hippocampal pyramids (Fig. 1, inset); silver grains are, however, not found in the nerve cells themselves but are concentrated in the surrounding neuropil. Activity is high also in the nucleus habenularis medialis. There is a moderate reaction in the cortex, especially in the vicinity of the interhemispherical fissure.

Both the molecular layer and the granular layer of the cerebellum contain numerous silver grains (Fig. 2). The localization pattern of the reaction does not, however, conform to any of the neural elements but resembles more closely the glial structure. The reaction is slightly stronger around Purkinje cells, which are themselves devoid of any reaction both in pre-fixed and in non-fixed specimens. On the other hand, in deep cerebellar nuclei (and also in some of the brain stem nuclei, for example the substantia nigra and nuclei pontis) silver grains are located within the cytoplasms of nerve cells. Weak or virtually no activity can be seen in the white matter, although the glial cells in the white matter react if higher doses of TSC are used or if the animals are killed after a shorter time interval.

The nerve cells in the spinal cord do not contain silver grains, but capsular and other glial cells in both the grey and white matter contain numerous grains (Fig. 3). According to Curtis[2], GABA is not involved in spinal inhibitory mechanisms.

The gross distribution of silver grains reduced by TSC-^{14}C is in accord with the distribution of GAD and GABA anticipated on the basis of earlier biochemical and pharmacological studies (cerebellar cortex[2], hippocampus[3], substantia nigra[6]). While the concentration of silver grains in nerve cells (dentate nucleus, substantia nigra) and in the neuropil surrounding them (hippocampus) is consistent with current views, it is striking that Purkinje cells do not exert any reaction. Proteins sensitive to TSC are undoubtedly more widespread than GAD, so the apparent localization of TSC in glial cells of the cerebellar cortex may be partly due to binding of the drug to other B_6-dependent enzymes. The possibility cannot be excluded, however, that, at least in this area, GABA is produced by glial cells, perhaps in order to be taken up by nerve terminals and/or nerve cells in a second step.

BERTALAN CSILLIK
ELIZABETH KNYIHÁR

Department of Anatomy,
University Medical School,
Szeged, Hungary.

Received August 12; revised October 6, 1969.

[1] Roberts, E., in *Structure and Function of Inhibitory Mechanisms*, 401 (Pergamon Press, Oxford, 1968).
[2] Curtis, D. R., in *Structure and Function of Inhibitory Mechanisms*, 429 (Pergamon Press, Oxford, 1968).
[3] Van Gelder, N. M., *J. Neurochem.*, 12, 231, 239 (1965).
[4] Killam, K. F., Dasgupta, S. R., and Killam, E. K., in *Inhibition in the Nervous System and Gamma-Aminobutyric Acid*, 302 (Pergamon Press, Oxford, 1960).
[5] Ostrovsky, K., and Barnard, E. A., *Exp. Cell Res.*, 25, 456 (1961).
[6] Albers, R. W., in *Inhibition in the Nervous System and Gamma-Aminobutyric Acid*, 196 (Pergamon Press, Oxford, 1960).

Natural Selection and the Concept of a Protein Space

SALISBURY[1] has argued that there is an apparent contradiction between two fundamental concepts of biology—the belief that the gene is a unique sequence of nucleotides whose function it is to determine the sequence of amino-acids in a protein, and the theory of evolution by natural selection. In brief, he calculated that the number of possible amino-acid sequences is greater by many orders of magnitude than the number of proteins which could have existed on Earth since the origin of life, and hence that functionally effective proteins have a vanishingly small chance of arising by mutation. Natural selection is

therefore ineffective because it lacks the essential raw material—favourable mutations.

I should like to look at the problem from a different point of view. I shall assume that mutations, while not random in a chemical sense, are random as far as their chances of improving the function of the corresponding proteins are concerned. I shall also assume that evolution has occurred either by the natural selection of favourable mutations or by the chance fixation by genetic drift of selectively neutral mutations. The justification for making these assumptions is that no sensible alternatives have been suggested and that no evidence exists at the moment to invalidate them. If these assumptions are true, what can we say about the frequency and distribution of amino-acid sequences which are functional, either as enzymes or in some other way?

The model of protein evolution I want to discuss is best understood by analogy with a popular word game. The object of the game is to pass from one word to another of the same length by changing one letter at a time, with the requirement that all the intermediate words are meaningful in the same language. Thus WORD can be converted into GENE in the minimum number of steps, as follows:

WORD WORE GORE GONE GENE

This is an analogue of evolution, in which the words represent proteins; the letters represent amino-acids; the alteration of a single letter corresponds to the simplest evolutionary step, the substitution of one amino-acid for another; and the requirement of meaning corresponds to the requirement that each unit step in evolution should be from one functional protein to another. The reason for the last requirement is as follows: suppose that a protein A B C D . . . exists, and that a protein a b C D . . . would be favoured by selection if it arose. Suppose further that the intermediates a B C D . . . and A b C D . . . are non-functional. These forms would arise by mutation, but would usually be eliminated by selection before a second mutation could occur. The double step from a b C D . . . to A B C D would thus be very unlikely to occur. Such double steps with unfavourable intermediates may occasionally occur, but are probably too rare to be important in evolution.

This is a model of the way in which one gene may change into another. An increase in the number of different genes in a single organism presumably occurs by the duplication of an already existing gene followed by divergence. If so, it remains true that new genes arise as modifications of pre-existing ones.

It follows that if evolution by natural selection is to occur, functional proteins must form a continuous network which can be traversed by unit mutational steps without passing through nonfunctional intermediates. In this respect, functional proteins resemble four-letter words in the English language, rather than eight-letter words, for the latter form a series of small isolated islands in a sea of nonsense sequences. Of course, this is not to deny the existence of isolated island proteins, analogous to the four-letter words ALSO and ALTO.

It is easy to state the condition which must be satisfied if meaningful proteins are to form a network. Let X be a meaningful protein. Let N be the number of proteins which can be derived from X by a unit mutational step, and f the fraction of these which are meaningful, in the sense of being as good as or better than X in some environment. Then, if $fN > 1$, meaningful proteins will form a network, and evolution by natural selection is possible. In estimating N it is necessary to distinguish two classes of mutations: (i) substitutions of single amino-acids, and additions or deletions of small numbers of amino-acids, making only a small change to the protein; and (ii) mutations producing a major change in amino-acid sequence, such as frame shifts and intramolecular inversions.

Mutations of the former type are much more likely to give rise to meaningful proteins than the latter. In the same way, a single random letter substitution in a meaningful word is more likely to give rise to a meaningful word than the simultaneous alteration of all the letters. Although frame shift mutations are known to occur, it is not clear whether they have ever been incorporated in evolution. It is therefore better to take N as the number of possible substitutions of single amino-acids. If all substitutions were possible in a single mutational step, N for a protein of 100 amino-acids would be 1,900. In practice the genetic code limits N to approximately 10^3.

Hence f must be greater than 1/1,000. It does not follow that the fraction of all possible sequences which are meaningful need be as high as 1/1,000. It is probably much lower. There is almost certainly a higher probability that a sequence will be meaningful if it is a neighbour of an existing functional protein than if it is selected at random. In fact, in treating N as the number of amino-acid substitutions rather than as the total number of possible mutational steps, it was in effect assumed that a random sequence has a negligibly small probability of being functional; this assumption will be confirmed if it turns out that frame shifts are rarely or never incorporated in evolution.

Suppose now that we imagine all possible amino-acid sequences to be arranged in a "protein space", so that two sequences are neighbours if one can be converted into another by a single amino-acid substitution. Then the requirement that $fN > 1$ requires that the "density" of functional proteins in certain regions of the space must be quite high—perhaps greater than 1/1,000. This agrees with Salisbury's conclusion that proteins, and hence the genes that determine them, cannot be as unique as all that. As a convinced Darwinist, I published[2] the conclusion that $fN > 1$ when little was known about the frequency of amino-acid substitutions in evolution. Since then evidence has accumulated (for a review, see King and Jukes[3]) that many substitutions are either selectively neutral or at least make comparatively minor changes in the function of proteins.

If $fN > 1$, no quantitative difficulty arises in explaining the evolution of proteins by natural selection. A difficulty nevertheless remains in explaining the origin of life—that is, in explaining the origin of the first functional proteins together with the genetic mechanism for producing them. If it were true that only a minute fraction of possible amino-acid sequences have even the slightest enzymatic activity, it would be difficult to understand how the first proteins arose. I do not want to discuss the problem of the origin of life, but only to point out that it is a quite different problem from that of the mechanism of evolution.

Some questions about molecular evolution can be formulated more clearly in terms of a protein space. For example: (i) Are all existing proteins part of the same continuous network, and if so, have they all been reached from a single starting point? Possible alternatives are that there are two or more distinct networks, or that there is one network with multiple starting points. (ii) How often, if ever, has evolution passed through a non-functional sequence? If so, has this been achieved by the random walk of genes rendered redundant by duplication, or by the chance concurrence of two or more mutations? (iii) What fraction of the functional network has already been explored in evolution? (iv) What fraction of potentially useful proteins are inaccessible?

JOHN MAYNARD SMITH

School of Biological Sciences,
University of Sussex.

Received November 7, 1969.

[1] Salisbury, F. B., *Nature*, 224, 342 (1969).
[2] Maynard Smith, J., in *The Scientist Speculates* (edit. by Good, I. J.) (Heinemann, London, 1961).
[3] King, J. L., and Jukes, T. H., *Science*, 164, 788 (1969).

Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitutions

JAMES U. BOWIE,* JOHN F. REIDHAAR-OLSON, WENDELL A. LIM, ROBERT T. SAUER

An amino acid sequence encodes a message that determines the shape and function of a protein. This message is highly degenerate in that many different sequences can code for proteins with essentially the same structure and activity. Comparison of different sequences with similar messages can reveal key features of the code and improve understanding of how a protein folds and how it performs its function.

THE GENOME IS MANIFEST LARGELY IN THE SET OF PROteins that it encodes. It is the ability of these proteins to fold into unique three-dimensional structures that allows them to function and carry out the instructions of the genome. Thus, comprehending the rules that relate amino acid sequence to structure is fundamental to an understanding of biological processes. Because an amino acid sequence contains all of the information necessary to determine the structure of a protein (*1*), it should be possible to predict structure from sequence, and subsequently to infer detailed aspects of function from the structure. However, both problems are extremely complex, and it seems unlikely that either will be solved in an exact manner in the near future. It may be possible to obtain approximate solutions by using experimental data to simplify the problem. In this article, we describe how an analysis of allowed amino acid substitutions in proteins can be used to reduce the complexity of sequences and reveal important aspects of structure and function.

Methods for Studying Tolerance to Sequence Variation

There are two main approaches to studying the tolerance of an amino acid sequence to change. The first method relies on the process of evolution, in which mutations are either accepted or rejected by natural selection. This method has been extremely powerful for proteins such as the globins or cytochromes, for which sequences from many different species are known (*2–7*). The second approach uses genetic methods to introduce amino acid changes at specific positions in a cloned gene and uses selections or screens to identify functional sequences. This approach has been used to great advantage for proteins that can be expressed in bacteria or yeast, where the appropriate genetic manipulations are possible (*3, 8–11*). The end results of both methods are lists of active sequences that can be compared and analyzed to identify sequence features that are essential for folding or function. If a particular property of a side chain, such as charge or size, is important at a given position, only side chains that have the required property will be allowed. Conversely, if the chemical identity of the side chain is unimportant, then many different substitutions will be permitted.

Studies in which these methods were used have revealed that proteins are surprisingly tolerant of amino acid substitutions (*2–4, 11*). For example, in studying the effects of approximately 1500 single amino acid substitutions at 142 positions in *lac* repressor, Miller and co-workers found that about one-half of all substitutions were phenotypically silent (*11*). At some positions, many different, nonconservative substitutions were allowed. Such residue positions play little or no role in structure and function. At other positions, no substitutions or only conservative substitutions were allowed. These residues are the most important for *lac* repressor activity.

What roles do invariant and conserved side chains play in proteins? Residues that are directly involved in protein functions such as binding or catalysis will certainly be among the most conserved. For example, replacing the Asp in the catalytic triad of trypsin with Asn results in a 10^4-fold reduction in activity (*12*). A similar loss of activity occurs in λ repressor when a DNA binding residue is changed from Asn to Asp (*13*). To carry out their function, however, these catalytic residues and binding residues must be precisely oriented in three dimensions. Consequently, mutations in residues that are required for structure formation or stability can also have dramatic effects on activity (*10, 14–16*). Hence, many of the residues that are conserved in sets of related sequences play structural roles.

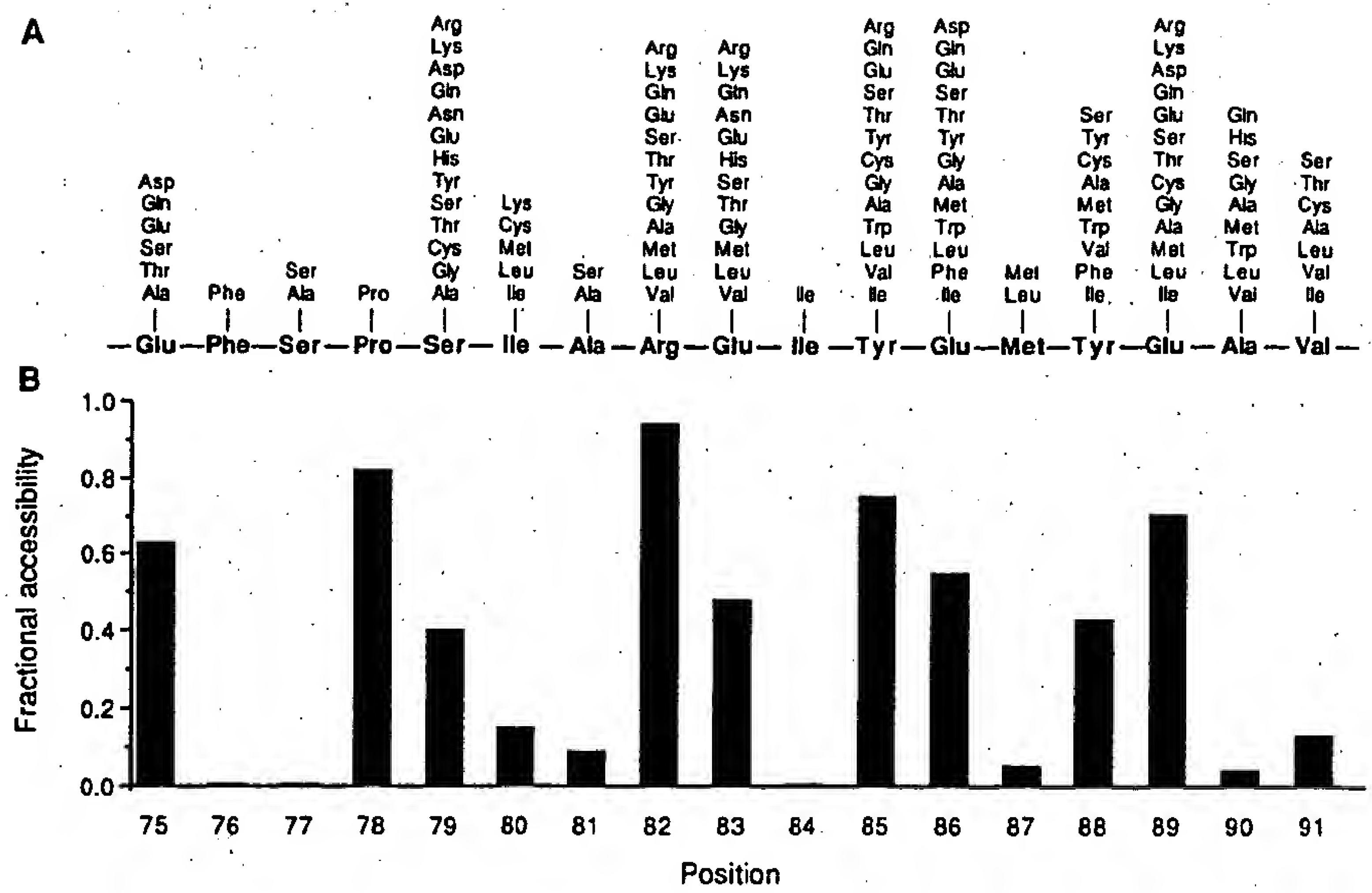
Substitutions at Surface and Buried Positions

In their initial comparisons of the globin sequences, Perutz and co-workers found that most buried residues require nonpolar side chains, whereas few features of surface side chains are generally conserved (*6*). Similar results have been seen for a number of protein families (*2, 4, 5, 7, 17, 18*). An example of the sequence tolerance at surface versus buried sites can be seen in Fig. 1, which shows the allowed substitutions in λ repressor at residue positions that are near the dimer interface but distant from the DNA binding surface of the protein (*9*). These substitutions were identified by a functional

The authors are in the Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139.

*Present address: Department of Chemistry and Biochemistry and the Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA 90024.

Fig. 1. (A) Amino acid substitutions allowed in a short region of λ repressor. The wild-type sequence is shown along the center line. The allowed substitutions shown above each position were identified by randomly mutating one to three codons at a time by using a cassette method and applying a functional selection (9). (B) The fractional solvent accessibility (42) of the wild-type side chain in the protein dimer (43) relative to the same atoms in an Ala-X-Ala model tripeptide.



selection after cassette mutagenesis. A histogram of side chain solvent accessibility in the crystal structure of the dimer is also shown in Fig. 1. At six positions, only the wild-type residue or relatively conservative substitutions are allowed. Five of these positions are buried in the protein. In contrast, most of the highly exposed positions tolerate a wide range of chemically different side chains, including hydrophilic and hydrophobic residues. Hence, it seems that most of the structural information in this region of the protein is carried by the residues that are solvent inaccessible.

Constraints on Core Sequences

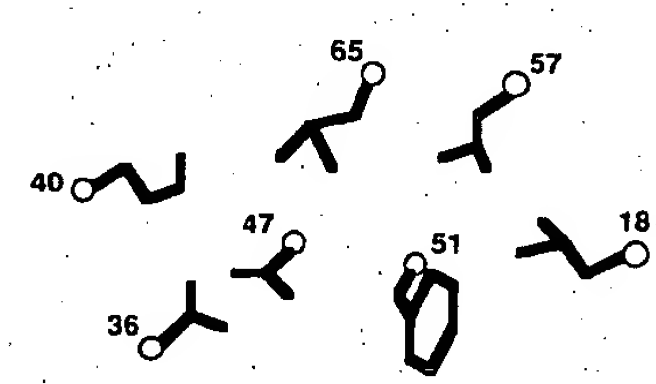
Because core residue positions appear to be extremely important for protein folding or stability, we must understand the factors that dictate whether a given core sequence will be acceptable. In general, only hydrophobic or neutral residues are tolerated at buried sites in proteins, undoubtedly because of the large favorable contribution of the hydrophobic effect to protein stability (19). For example, Fig. 2 shows the results of genetic studies used to investigate the substitutions allowed at residue positions that form the hydrophobic core of the NH_2-terminal domain of λ repressor (20). The acceptable core sequences are composed almost exclusively of Ala, Cys, Thr, Val, Ile, Leu, Met, and Phe. The acceptability of many different residues at each core position presumably reflects the fact that the hydrophobic effect, unlike hydrogen bonding, does not depend on specific residue pairings. Although it is possible to imagine a hypothetical core structure that is stabilized exclusively by residues forming hydrogen bonds and salt bridges, such a core would probably be difficult to construct because hydrogen bonds require pairing of donors and acceptors in an exact geometry. Thus the repertoire of possible structures that use a polar core would probably be extremely limited (21). Polar and charged residues are occasionally found in the cores of proteins, but only at positions where their hydrogen bonding needs can be satisfied (22).

The cores of most proteins are quite closely packed (23), but some volume changes are acceptable. In λ repressor, the overall core volume of acceptable sequences can vary by about 10%. Changes at individual sites, however, can be considerably larger. For example, as shown in Fig. 2, both Phe and Ala are allowed at the same core position in the appropriate sequence contexts. Large volume changes at individual buried sites have also been observed in phylogenetic studies, where it has been noted that the size decreases and increases at interacting residues are not necessarily related in a simple complementary fashion (5, 7, 17). Rather, local volume changes are accommodated by conformational changes in nearby side chains and by a variety of backbone movements.

The Informational Importance of the Core

With occasional exceptions, the core must remain hydrophobic and maintain a reasonable packing density. However, since the core is composed of side chains that can assume only a limited number of conformations (24), efficient packing must be maintained without steric clashes. How important are hydrophobicity, volume, and steric complementarity in determining whether a given sequence can form an acceptable core? Each factor is essential in a physical sense, as a stable core is probably unable to tolerate unsatisfied hydrogen bonding groups, large holes, or steric overlaps (25). However, in an informational sense, these factors are not equivalent. For example, in experiments in which three core residues of λ repressor were mutated simultaneously, volume was a relatively unimportant informational constraint because three-quarters of all possible combinations of the 20 naturally occurring amino acids had volumes within the range tolerated in the core, and yet most of these sequences were unacceptable (20). In contrast, of the sequences that contained only

Fig. 2. Amino acid substitutions allowed in the core of λ repressor. The wild-type side chains are shown pictorially in the approximate orientation seen in the crystal structure (43). The lists of allowed substitutions at each position are shown below the wild-type side chains. These substitutions were identified by randomly mutating one to four residues at a time by using a cassette method and applying a functional selection (20). Not all substitutions are allowed in every sequence background.



40	36	47	65	51	57	18
Ala	Ala	Ala	Ala	Cys	Cys	Ala
Cys	Cys	Cys	Ser	Val	Pro	Cys
Val	Thr	Thr	Cys	Ile	Val	Val
Ile	Val	Val	Thr	Leu	Ile	Ile
Leu	Ile	Ile	Val	Met	Leu	Leu
Met	Leu	Leu	Ile	Phe	Met	Met
Phe	Met	Met	Leu		Phe	
	Phe	Phe	Met			
			Phe			

the appropriate hydrophobic residues, a significant fraction were acceptable. Hence, the hydrophobicity of a sequence contains more information about its potential acceptability in the core than does the total side chain volume. Steric compatibility was intermediate between volume and hydrophobicity in informational importance.

The Informational Importance of Surface Sites

We have noted that many surface sites can tolerate a wide variety of side chains, including hydrophilic and hydrophobic residues. This result might be taken to indicate that surface positions contain little structural information. However, Bashford *et al.*, in an extensive analysis of globin sequences (*4*), found a strong bias against large hydrophobic residues at many surface positions. At one level, this may reflect constraints imposed by protein solubility, because large patches of hydrophobic surface residues would presumably lead to aggregation. At a more fundamental level, protein folding requires a partitioning between surface and buried positions. Consequently, to achieve a unique native state without significant competition from other conformations, it may be important that some sites have a decided preference for exterior rather than interior positions. As a result, many surface sites can accept hydrophobic residues individually, but the surface as a whole can probably tolerate only a moderate number of hydrophobic side chains.

Identification of Residue Roles from Sets of Sequences

Often, a protein of interest is a member of a family of related sequences. What can we infer from the pattern of allowed substitutions at positions in sets of aligned sequences generated by genetic or phylogenetic methods? Residue positions that can accept a number of different side chains, including charged and highly polar residues, are almost certain to be on the protein surface. Residue positions that remain hydrophobic, whether variable or not, are likely to be buried within the structure. In Fig. 3, those residue positions in λ repressor that can accept hydrophilic side chains are shown in orange and those that cannot accept hydrophilic side chains are shown in green. The obligate hydrophobic positions define the core of the structure, whereas positions that can accept hydrophilic side chains define the surface.

Functionally important residues should be conserved in sets of active sequences, but it is not possible to decide whether a side chain is functionally or structurally important just because it is invariant or conserved. To make this distinction requires an independent assay of protein folding. The ability of a mutant protein to maintain a stably folded structure can often be measured by biophysical techniques, by susceptibility to intracellular proteolysis (*26*), or by binding to antibodies specific for the native structure (*27, 28*). In the latter cases, it is possible to screen proteins in mutated clones for the ability to fold even if these proteins are inactive. Sets of sequences that allow formation of a stable structure can then be compared to the sets that allow both folding and function, with the active site or binding residues being those that are variable in the set of stable proteins but invariant in the set of functional proteins. The DNA-binding residues of Arc repressor were identified by this method (*8*). The receptor-binding residues of human growth hormone were also identified by comparing the stabilities and activities of a set of mutant sequences (*28*). However, in this case, the mutants were generated as hybrid sequences between growth hormone and related hormones with different binding specificities.

Implications for Structure Prediction

At present, the only reliable method for predicting a low-resolution tertiary structure of a new protein is by identifying sequence similarity to a protein whose structure is already known (*29, 30*). However, it is often difficult to align sequences as the level of sequence similarity decreases, and it is sometimes impossible to detect statistically significant sequence similarity between distantly related proteins. Because the number of known sequences is far greater than the number of known structures, it would be advantageous to increase the reach of the available structural information by improving methods for detecting distant sequence relations and for subsequently aligning these sequences based on structural principles. In a normal homology search, the sequence database is scanned with a single test sequence, and every residue must be weighted equally. However, some residues are more important than others and should be weighted accordingly. Moreover, certain regions of the protein are more likely to contain gaps than others. Both kinds of information can be obtained from sequence sets, and several techniques have

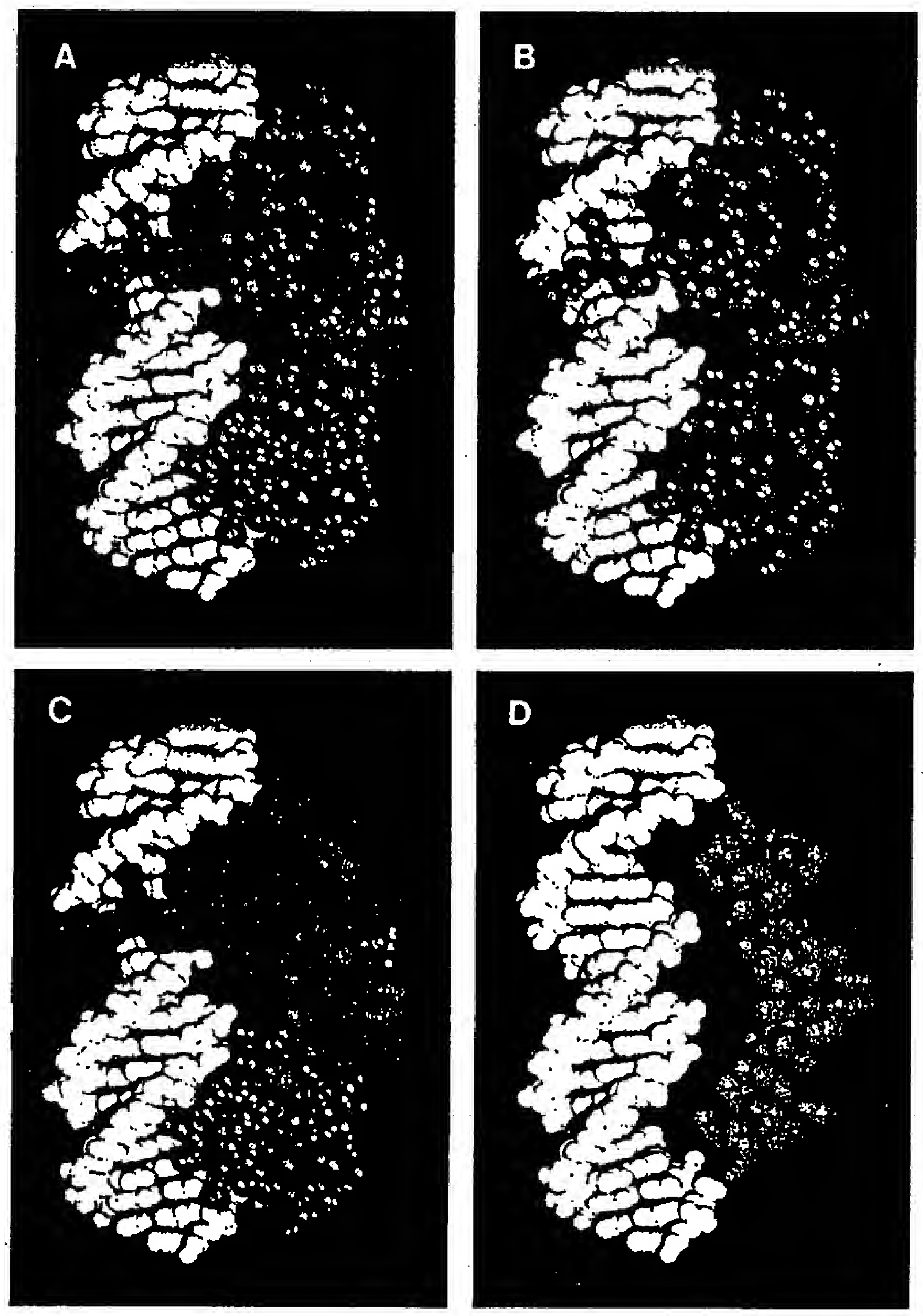


Fig. 3. Tolerance of positions in the NH_2-terminal domain of λ repressor to hydrophilic side chains. The complex (*43*) of the repressor dimer (blue) and operator DNA (white) is shown. In (**A**), positions that can tolerate hydrophilic side chains are shown in orange. The same side chains are shown in (**B**) without the remaining protein atoms. In (**C**), positions that require hydrophobic or neutral side chains are shown in green. These side chains are shown in (**D**) without the remaining protein atoms. About three-fourths of the 92 side chains in the NH_2-terminal domain are included in both (B) and (D). The remaining positions have not been tested. Data are from (*9, 14, 20, 27, 44*).

been used to combine such information into more appropriately weighted sequence searches and alignments (*31*). These methods were used to align the sequences of retroviral proteases with aspartic proteases, which in turn allowed construction of a three-dimensional model for the protease of human immunodeficiency virus type 1 (*29*). Comparison with the recently determined crystal structure of this protein revealed reasonable agreement in many areas of the predicted structure (*32*).

The structural information at most surface sites is highly degenerate. Except for functionally important residues, exterior positions seem to be important chiefly in maintaining a reasonably polar surface. The information contained in buried residues is also degenerate, the main requirement being that these residues remain hydrophobic. Thus, at its most basic level, the key structural message in an amino acid sequence may reside in its specific pattern of hydrophobic and hydrophilic residues. This is meant in an informational sense. Clearly, the precise structure and stability of a protein depends on a large number of detailed interactions. It is possible, however, that structural prediction at a more primitive level can be accomplished by concentrating on the most basic informational aspects of an amino acid sequence. For example, amphipathic patterns can be extracted from aligned sets of sequences and used, in some cases, to identify secondary structures.

If a region of secondary structure is packed against the hydrophobic core, a pattern of hydrophobic residues reflecting the periodicity of the secondary structure is expected (*33, 34*). These patterns can be obscured in individual sequences by hydrophobic residues on the protein surface. It is rare, however, for a surface position to remain hydrophobic over the course of evolution. Consequently, the amphipathic patterns expected for simple secondary structures can be much clearer in a set of related sequences (*6*). This principle is illustrated in Fig. 4, which shows helical hydrophobic moment plots for the Antennapedia homeodomain sequence (Fig. 4A) and for a composite sequence derived from a set of homologous homeodomain proteins (Fig. 4B) (*35*). The hydrophobic moment is a simple measure of the degree of amphipathic character of a sequence in a given secondary structure (*34*). The amphipathic character of the three α-helical regions in the Antennapedia protein (*36*) is clearly revealed only by the analysis of the combined set of homeodomain sequences. The secondary structure of Arc repressor, a small DNA-binding protein, was recently predicted by a similar method (*8*) and confirmed by nuclear magnetic resonance studies (*37*).

The specific pattern of hydrophobic and hydrophilic residues in an amino acid sequence must limit the number of different structures a given sequence can adopt and may indeed define its overall fold. If this is true, then the arrangement of hydrophobic and hydrophilic residues should be a characteristic feature of a particular fold. Sweet and Eisenberg have shown that the correlation of the pattern of hydrophobicity between two protein sequences is a good criterion for their structural relatedness (*38*). In addition, several studies indicate that patterns of obligatory hydrophobic positions identified from aligned sequences are distinctive features of sequences that adopt the same structure (*4, 29, 38, 39*). Thus, the order of hydrophobic and hydrophilic residues in a sequence may actually be sufficient information to determine the basic folding pattern of a protein sequence.

Although the pattern of sequence hydrophobicity may be a characteristic feature of a particular fold, it is not yet clear how such patterns could be used for prediction of structure de novo. It is important to understand how patterns in sequence space can be related to structures in conformation space. Lau and Dill have approached this problem by studying the properties of simple sequences composed only of H (hydrophobic) and P (polar) groups on two-dimensional lattices (*40*). An example of such a representation is shown in Fig. 5. Residues adjacent in the sequence must occupy adjacent squares on the lattice, and two residues cannot occupy the same space. Free energies of particular conformations are evaluated with a single term, an attraction of H groups. By considering chains of ten residues, an exhaustive conformational search for all 1024 possible sequences of H and P residues was possible. For longer sequences only a representative fraction of the allowed sequence or conformation space could be explored. The significant results were as follows: (i) not all sequences can fold into a "native" structure and only a few sequences form a unique native structure; (ii) the probability that a sequence will adopt a unique native structure increases with chain length; and (iii) the native states are compact, contain a hydrophobic core surrounded by polar residues, and contain significant secondary structure. Although the gap between these two-dimensional simulations and three-dimensional structures is large, the use of simple rules and sequence representations yields results similar to those expected for real proteins. Three-dimensional lattice methods are also beginning to be developed and evaluated (*41*).

Summary

There is more information in a set of related sequences than in a single sequence. A number of practical applications arise from an analysis of the tolerance of residue positions to change. First, such information permits the evaluation of a residue's importance to the function and stability of a protein. This ability to identify the essential elements of a protein sequence may improve our understanding of the determinants of protein folding and stability as well as protein function. Second, patterns of tolerance to amino acid substitutions of varying hydrophilicity can help to identify residues likely to be buried in a protein structure and those likely to occupy

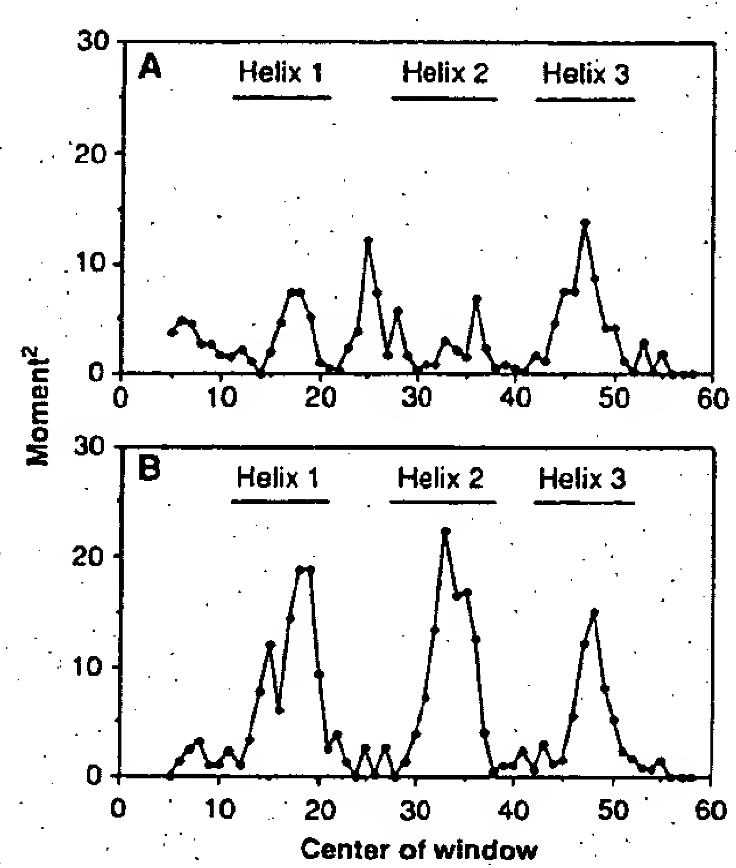


Fig. 4. Helical hydrophobic moments calculated by using (**A**) the Antennapedia homeodomain sequence or (**B**) a set of 39 aligned homeodomain sequences (*35*). The bars indicate the extent of the helical regions identified in nuclear magnetic resonance studies of the Antennapedia homeodomain (*36*). To determine hydrophobic moments, residues were assigned to one of three groups: H1 (high hydrophobicity = Trp, Ile, Phe, Leu, Met, Val, or Cys); H2 (medium hydrophobicity = Tyr, Pro, Ala, Thr, His, Gly, or Ser); and H3 (low hydrophobicity = Gln, Asn, Glu, Asp, Lys, or Arg). For the aligned homeodomain sequences, the residues at each position were sorted by their hydrophobicity by using the scale of Fauchere and Pliska (*45*). Arg and Lys were not counted unless no other residue was found at the position, because they contain long aliphatic side chains and can thereby substitute for nonpolar residues at some buried sites. To account for possible sequence errors and rare exceptions, the most hydrophilic residue allowed at each position was discarded unless it was observed twice. The second most hydrophilic residue was then chosen to represent the hydrophobicity of each position. An eight-residue window was used and the vectors projected radially every 100°. The vector magnitudes were assigned a value of 1, 0, or −1 for positions where the hydrophobicity group was H1, H2, or H3, respectively.

P H P P H P H P H H H P P H

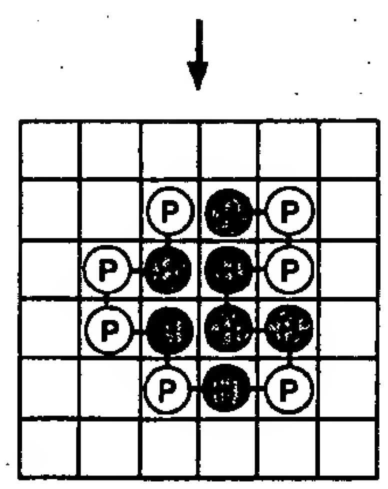


Fig. 5. A representation of one compact conformation for a particular sequence of H and P residues on a two-dimensional square lattice. [Adapted from (*40*), with permission of the American Chemical Society]

surface positions. The amphipathic patterns that emerge can be used to identify probable regions of secondary structure. Third, incorporating a knowledge of allowed substitutions can improve the ability to detect and align distantly related proteins because the essential residues can be given prominence in the alignment scoring.

As more sequences are determined, it becomes increasingly likely that a protein of interest is a member of a family of related sequences. If this is not the case, it is now possible to use genetic methods to generate lists of allowed amino acid substitutions. Consequently, at least in the short term, it may not be necessary to solve the folding problem for individual protein sequences. Instead, information from sequence sets could be used. Perhaps by simplifying sequence space through the identification of key residues, and by simplifying conformation space as in the lattice methods, it will be possible to develop algorithms to generate a limited number of trial structures. These trial structures could then, in turn, be evaluated by further experiments and more sophisticated energy calculations.

REFERENCES AND NOTES

1. C. J. Epstein, R. F. Goldberger, C. B. Anfinsen, *Cold Spring Harbor Symp. Quant. Biol.* **28**, 439 (1963); C. B. Anfinsen, *Science* **181**, 223 (1973).
2. R. E. Dickerson, *Sci. Am.* **242**, 136 (March 1980).
3. M. D. Hampsey, G. Das, F. Sherman, *FEBS Lett.* **231**, 275 (1988).
4. D. Bashford, C. Chothia, A. M. Lesk, *J. Mol. Biol.* **196**, 199 (1987).
5. A. M. Lesk and C. Chothia, *ibid.* **136**, 225 (1980).
6. M. F. Perutz, J. C. Kendrew, H. C. Watson, *ibid.* **13**, 669 (1965).
7. C. Chothia and A. M. Lesk, *Cold Spring Harbor Symp. Quant. Biol.* **52**, 399 (1965).
8. J. U. Bowie and R. T. Sauer, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2152 (1989).
9. J. F. Reidhaar-Olson and R. T. Sauer, *Science* **241**, 53 (1988); *Proteins Struct. Funct. Genet.*, in press.
10. D. Shortle, *J. Biol. Chem.* **264**, 5315 (1989).
11. J. H. Miller *et al.*, *J. Mol. Biol.* **131**, 191 (1979).
12. S. Sprang *et al.*, *Science* **237**, 905 (1987); C. S. Craik, S. Roczniak, C. Largman, W. J. Rutter, *ibid.*, p. 909.
13. H. C. M. Nelson and R. T. Sauer, *J. Mol. Biol.* **192**, 27 (1986).
14. M. H. Hecht, J. M. Sturtevant, R. T. Sauer, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 5685 (1984).
15. T. Alber, D. Sun, J. A. Nye, D. C. Muchmore, B. W. Matthews, *Biochemistry* **26**, 3754 (1987).
16. D. Shortle and A. K. Meeker, *Proteins Struct. Funct. Genet.* **1**, 81 (1986).
17. A. M. Lesk and C. Chothia, *J. Mol. Biol.* **160**, 325 (1982).
18. W. R. Taylor, *ibid.* **188**, 233 (1986).
19. W. Kauzmann, *Adv. Protein Chem.* **14**, 1 (1959); R. L. Baldwin, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 8069 (1986).
20. W. A. Lim and R. T. Sauer, *Nature* **339**, 31 (1989); in preparation.
21. Lesk and Chothia (*5*) have argued that a protein core composed solely of hydrogen-bonded residues would also be inviable on evolutionary grounds, as a mutational change in one core residue would require compensating changes in any interacting residue or residues to maintain a stable structure.
22. T. M. Gray and B. W. Matthews, *J. Mol. Biol.* **175**, 75 (1984); E. N. Baker and R. E. Hubbard, *Prog. Biophys. Mol. Biol.* **44**, 97 (1984).
23. F. M. Richards, *J. Mol. Biol.* **82**, 1 (1974).
24. J. W. Ponder and F. M. Richards, *ibid.* **193**, 775 (1987).
25. J. T. Kellis, Jr., K. Nyberg, A. R. Fersht, *Biochemistry* **28**, 4914 (1989); W. S. Sandberg and T. C. Terwilliger, *Science* **245**, 54 (1989).
26. A. A. Pakula and R. T. Sauer, *Proteins Struct. Funct. Genet.* **5**, 202 (1989).
27. B. C. Cunningham and J. A. Wells, *Science* **244**, 1081 (1989); R. M. Breyer and R. T. Sauer, *J. Biol. Chem.* **264**, 13348 (1989).
28. B. C. Cunningham, P. Jhurani, P. Ng, J. A. Wells, *Science* **243**, 1330 (1989).
29. L. H. Pearl and W. R. Taylor, *Nature* **329**, 351 (1987).
30. W. J. Brown *et al.*, *J. Mol. Biol.* **42**, 65 (1969); J. Greer, *ibid.* **153**, 1027 (1981); J. M. Berg, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 99 (1988).
31. W. R. Taylor, *Protein Eng.* **2**, 77 (1988).
32. M. A. Navia *et al.*, *Nature* **337**, 615 (1989).
33. M. Schiffer and A. B. Edmundson, *Biophys. J.* **7**, 121 (1967); V. I. Lim, *J. Mol. Biol.* **88**, 857 (1974); *ibid.*, p. 873.
34. D. Eisenberg, R. M. Weiss, T. C. Terwilliger, *Nature* **299**, 371 (1982); D. Eisenberg, D. Schwarz, M. Komaromy, R. Wall, *J. Mol. Biol.* **179**, 125 (1984); D. Eisenberg, R. M. Weiss, T. C. Terwilliger, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 140 (1984).
35. T. R. Burglin, *Cell* **53**, 339 (1988).
36. G. Otting *et al.*, *EMBO J.* **7**, 4305 (1988).
37. J. N. Breg, R. Boelens, A. V. E. George, R. Kaptein, *Biochemistry* **28**, 9826 (1989); M. G. Zagorski, J. U. Bowie, A. K. Vershon, R. T. Sauer, D. J. Patel, *ibid.*, p. 9813.
38. R. M. Sweet and D. Eisenberg, *J. Mol. Biol.* **171**, 479 (1983).
39. J. U. Bowie, N. D. Clarke, C. O. Pabo, R. T. Sauer, *Proteins Struct. Funct. Genet.*, in preparation.
40. K. F. Lau and K. A. Dill, *Macromolecules* **22**, 3986 (1989).
41. A. Sikorski and J. Skolnick, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2668 (1989); A. Kolinski, J. Skolnick, R. Yaris, *Biopolymers* **26**, 937 (1987); D. G. Covell and R. L. Jernigan, *Biochemistry*, in press.
42. B. Lee and F. M. Richards, *J. Mol. Biol.* **55**, 379 (1971).
43. S. R. Jordan and C. O. Pabo, *Science* **242**, 893 (1988).
44. R. M. Breyer, thesis, Massachusetts Institute of Technology, Cambridge (1988).
45. J.-L. Fauchere and V. Pliska, *Eur. J. Med. Chem.-Chim. Ther.* **18**, 369 (1983).
46. We thank C. O. Pabo and S. Jordan for coordinates of the NH_2-terminal domain of λ repressor and its operator complex. We also thank P. Schimmel for the use of his graphics system and J. Burnbaum and C. Francklyn for assistance. Supported in part by NIH grant AI-15706 and predoctoral grants from NSF (J.R.-O.) and Howard Hughes Medical Institute (W.A.L.).

Table 1
Efficiency of suppression and specificity of insertion of tRNA suppressors

Suppressor	Gene	Amino acid inserted	Efficiency of suppression at UAG codons (%)
Su1	*supD*	Serine	6·54
Su2-89	*supE*	Glutamine	32–60
Su3	*supF*	Tyrosine	11–100
Su5	*supG*	Lysine	0·6–3; 5–29
Su6	*supP*	Leucine	30–100
$tRNA^{Ala}_{CUA}$	synthetic	Alanine	8–83
$tRNA^{Cys}_{CUA}$	synthetic	Cysteine	17–54
$tRNA^{GluA}_{CUA}$	synthetic	Glutamic Acid (80%) Glutamine (20%)	8–100
$tRNA^{Gly\,1}_{CUA}$	synthetic	Glycine	24–100
$tRNA^{HisA}_{CUA}$	synthetic	Histidine	16–100
$tRNA^{Lys}_{CUA}$	synthetic	Lysine	9–29
$tRNA^{Phe}_{CUA}$	synthetic	Phenylalanine	48–100
$tRNA^{ProH}_{CUA}$	synthetic	Proline	9—60
FTOIRΔ26	synthetic	Arginine	4–28; 4–47

Data has been compiled from previous work on naturally occurring (Miller & Albertini, 1983) and synthetic (Kleina *et al.*, 1990) suppressors. The efficiency of suppression is expressed as a range of the highest and lowest values obtained in different contexts. The Su2-89 suppressor is from Bradley *et al.* (1981), and the FTOR1 26 suppressor from McClain & Foss (1988).

Table 2
Phenotypes resulting from different lacI *mutations*

Altered repressor function/property	Phenotypic symbol	Mutation dominant/recessive (d) (r)
DNA binding	I^-	d
Folding	I^-	r
Aggregation	I^-	r
Inducer binding	I^S	d
Allosteric transition	I^S	d
Tight binding to DNA	I^S, I^r	
Reversed allosteric transition	I^r, I^{rc}	
Stability increased by inducer binding	I^r	

Some of the altered repressors which have been characterized are shown. The I^- designation is for different defective repressors which can no longer block transcription of the *lac* genes. The I^S symbol denotes repressors that bind to operator but are not induced by IPTG. Several types of altered repressors display this characteristic, including those that bind operator more tightly. This latter class also displays a partial reverse induction profile (in which repression increases with increasing IPTG concentration), which is designated as I^r (Chamness & Wilson, 1970). I^r of I^{rc} (reverse curve; Myers & Sadler, 1971) repressors may also result from altered allosteric transition or from a stabilization of repressor by inducer binding. Temperature-sensitive derivatives of each type of repressor can also occur. For a more detailed description, see Kleina & Miller (1990), Mueller-Hill (1975) or Miller (1978).

amino acids at positions in a protein corresponding to a UAG site (Normanly *et al.*, 1990; Kleina *et al.*, 1990; McClain & Foss, 1988), and an additional amino acid at UGA sites. Table 1 depicts the amino acids which can be added by suppression of nonsense mutations at reasonable efficiencies.

The *E. coli lac* repressor is a tetrameric protein composed of four identical subunits, each consisting of 360 amino acids (Mueller-Hill, 1975; Farabaugh, 1978). The repressor consists of several domains. The amino-terminal 59 amino acids contain the DNA and operator binding sites, and the remaining portion (60 to 329) of the molecule contains the binding site for inducer and the dimer association sites (Mueller-Hill, 1975; Platt *et al.*, 1973; Ogata & Gilbert, 1978; Schmitz *et al.*, 1976; Miller & Schmeissner, 1979). The carboxyl-terminal 30–31 residues (330 to 360) are not required for the formation of active dimers, but are required for the dimer to tetramer transition (Platt *et al.*, 1970; Alberti *et al.*, 1991), and appear to contain a leucine zipper (Alberti *et al.*, 1991; Chakerian *et al.*, 1991). Numerous investigations have identified several classes of altered *lac* repressors (Mueller-Hill, 1975; Schmitz *et al.*, 1976; Miller & Schmeissner, 1979; Chamness & Willson, 1970; Miller, 1978) defective in one of these functions. Table 2 displays some of the phenotypes resulting from different *lacI* mutations.

We previously described the use of nonsense suppressors on a set of 141 nonsense mutations in the *lacI* gene (Kleina & Miller, 1990), which encodes the 360 amino acid *lac* repressor monomer. This work resulted in the generation of close to 1600 amino acid replacements. Using site-directed mutagenesis, we have constructed an additional 188 amber sites in the *lacI* gene. Here, we describe the effects of suppressing all of these mutations, spanning residues 2 to 329, with each of the characterized amber suppressors. (Amino-terminal fragments containing > 330 residues are partially or fully active *in vivo* and were not studied.) The replacement of 12–13 amino acids at each of the 328 amber sites results in over 4000 altered *lac* repressors, and allows a virtual genetic image reconstruction of the functional regions of the protein. We compare the results reported here from those of other large collections of altered proteins, including the pioneering studies of Perutz and co-workers on the hemoglobins and myoglobins (Perutz, 1965; Perutz *et al.*, 1965; Perutz & Lehmann, 1968), studies of HIV-1 protease (Loeb *et al.*, 1989), an N-terminal fragment of the lambda repressor (Bowie *et al.*, 1990), and T4 lysozyme (Rennell *et al.*, 1991). Additionally, we compare the substitutional tolerance of individual sites in the repressor with the evolutionary variability of those sites in homologous proteins.

2. Materials and Methods

(a) *Genetic methodology*

All genetic methods and assays were carried out as described by Miller (1972) and Kleina & Miller (1990). Mutants were constructed as described previously (Kleina & Miller, 1990), except that in many cases we crossed mutations directly from f1 onto the F′*lacpro* episome, without first cloning them onto a plasmid.

(b) *Correlation of tolerance and conservation*

The correlation coefficient, C_s (Schulz & Schirmer, 1979) is defined as:

$$C_s = \frac{(w)(x)-(y)(z)}{\sqrt{(x+y)(x+z)(w+y)(w+z)}}$$

where, as a percentage, or fraction: w is positive correct prediction, x is negative correction prediction, y is underprediction, z is overprediction, and where $w+x+y+z=1$. Here, w is conserved, intolerant, x is not conserved, tolerant, y is predicted conserved, but not: i.e. nonconserved, intolerant, z is conserved, tolerant;

$$w = 97/328 = 0{\cdot}30; \quad x = 163/328 = 0{\cdot}50;$$
$$y = 37/328 = 0{\cdot}11; \quad z = 31/328 = 0{\cdot}09.$$
$$C_s = \frac{0{\cdot}15-0{\cdot}0099}{\sqrt{(0{\cdot}61)(0{\cdot}59)(0{\cdot}41)(0{\cdot}39)}}.$$

$C_s = 0{\cdot}58$, which is highly significant (Matthews, 1975).

(c) *Scoring for + and − among altered repressors*

In order for a replacement to score as +, it must result in a repressor with greater than 8 to 10% activity. Although the version of the Su2 glutamine-inserting suppressor (Su2–89) that we are using operates with high efficiency, the normal Su2 operates at a very low efficiency at certain amber sites. When coupled with the amber site corresponding to residue 117, glutamine is inserted at only 0·8 to 1·0% efficiency (Miller & Albertini, 1983; J. H. Miller, unpublished results). Since the repressor is overproduced 10-fold (Mueller-Hill *et al.*, 1968) in the strain employed, that means that glutamine is being inserted at 8 to 10% of the level of a typical wild-type strain. Even though glutamine is the wild-type amino acid at position 117, repression is not fully restored with the normal Su2 suppressor. Therefore, replacements must result in greater than 8 to 10% activity in order to be scored as a full +.

3. Results

(a) *Effects of amino acid replacements*

Figure 1 compiles all of the accumulated data for the 13 × 328 amino acid replacement matrix in the *lac* repressor. This Figure shows the fractional tolerance to substitution for each residue in the protein, without considering each substitution in detail. More quantitative data on the effects of specific replacement are given by Kleina *et al.* (1990), and additional quantitative data will be published elsewhere. On examining the linear array of sites in the protein, several features strike the eye. It is clear that large portions of the protein (60%) are tolerant to substitutions. However, two subregions sensitive to replacements are evident. The first, consisting of the amino-terminal 59 residues, has previously been shown to be involved in operator and DNA binding (Mueller-Hill, 1975; Platt *et al.*, 1973; Ogata & Gilbert, 1978), and contains the helix-turn-helix DNA binding motif seen in many regulatory proteins (McKay & Steitz, 1981; Kaptein *et al.*, 1985). Analysis of the properties of sites in this region shows that the sensitivity to amino acid exchanges can be correlated with the three-dimensional structure of this region as determined by NMR (Kaptein *et al.*, 1985), and depicted in Figure 2. Three helices are defined in the first 51 amino acids. Helix II is the recognition helix, including amino acids 16 to 23 (Lehming *et al.*, 1987, 1988). It can be seen that all amino acids in this helix are extremely sensitive to substitutions. It is particularly striking that for helix III, the residues whose side-chains point inwards making internal contacts (e.g. Val38, Ala41, Met42 and Tyr47) are sensitive to replacements, while those facing the exterior (e.g. Glu36, Gln39, Ala43, Glu44 and Asn46) are very tolerant of substitutions. A similar pattern is seen for helix I, with Val9, Ala10 and Ala13 intolerant, while other residues pointing outward are more substitutable. Thr5 is highly intolerant, but is believed to be in a close non-specific contact with the operator DNA (Kaptein *et al.*, 1985). The residues forming the loops between the helices are for the most part tolerant to substitutions.

The second region containing many sites intolerant to substitutions extends from amino acids 239 to 241 until amino acids 289 to 292. The experimental identification of this region as a crucial portion of the repressor has emerged from this study. We have carried out a computer search in GenBank and other data bases for proteins with significant homologies to *E. coli lac* repressor, and have found that the region of the repressor, aside from the amino-terminal DNA binding domain, which shares the greatest homology with other proteins is in fact in the region from residues 239 to 289 (see below, and Figures 3–4).

Figure 1 also reveals stretches of amino acids which appear to be almost completely tolerant to substitutions. For instance, very few substitutions between residues 100 to 112, 129 to 145, 151 to 160, 206 to 217 or 305 to 318 appear to be deleterious. Throughout the protein, it appears as if long stretches of tolerant residues separate one or small clusters of sites which are sensitive to substitutions. In most of these cases the sensitive residues form hydrophobic clusters which are probably buried or partly buried in the interior of the protein.

(b) *Inducer binding*

Replacements resulting in the I^S phenotype, inability to respond to inducer, are scored in the upper portion of Figure 1. Since the vast majority of sites in the repressor are not involved in inducer binding, only those sites where an effect occurs are shown. I^S mutations affect binding of the inducer molecule and/or the allosteric transition that reduces binding affinity to operator. The strongest effects are shown as darkened boxes (see legend to Figure 1). In addition to a large array of I^S sites from amino acids 66 to 99, and scattered sites over the next 90 residues, the remainder of the protein contains distinct clusters of I^S sites at regularly spaced intervals (see lower portion of Figure 1 and Miller *et al.*, 1979). Exactly how these sites are arranged in three dimensions will have to await the elucidation of the repressor structure.

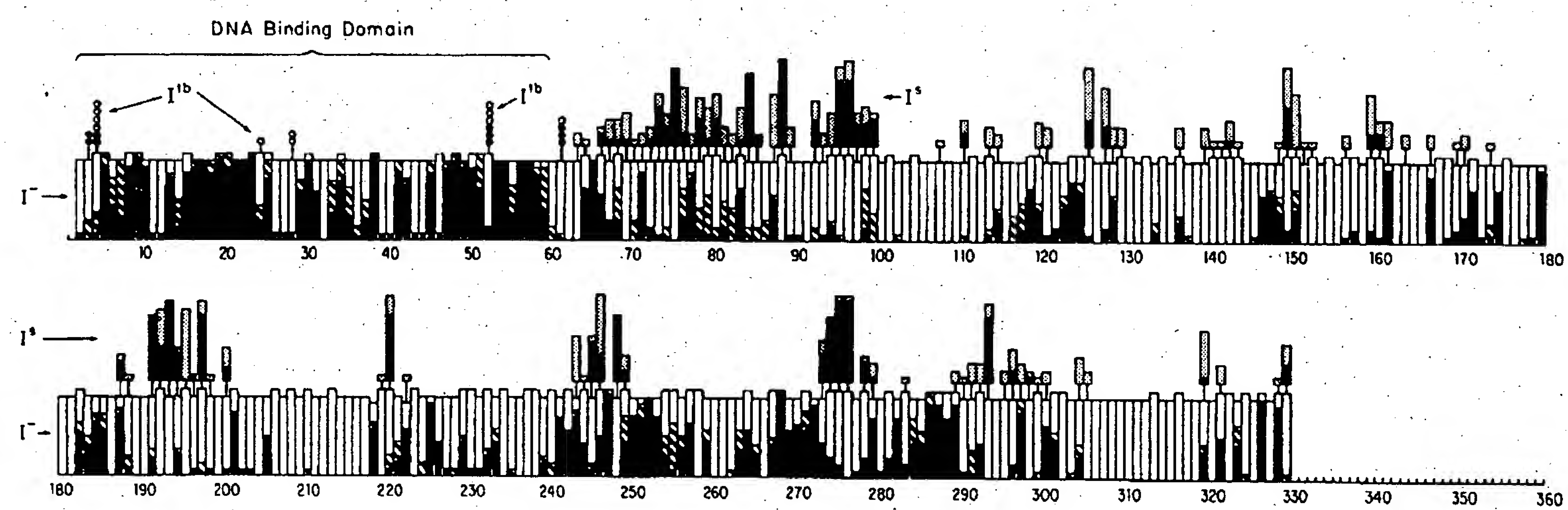


Figure 1. Effects of amino acid replacements in the *lac* repressor. The consequences of over 4000 single amino acid replacements are depicted here. The repressor protein is shown as an array of 360 sites (amino acids). An amber (UAG) mutation has been constructed at each position in the *lacI* gene, one mutation per construct, corresponding to residues 2 to 329 of the *lac* repressor (see Materials and Methods). At each position between residues 2 and 329, 12–13 amino acids can be exchanged through nonsense suppression and a bar has been drawn on the linear axis of the repressor polypeptide. The height of the bar represents the number of exchanges made (12 or 13). Each bar is divided into invisible segments, the number of segments corresponding to the number of replacements made. When a replacement generates the I^- phenotype, a corresponding portion of the bar is filled in. If the replacement does not alter the I^- phenotype, then the segment is left open or blank. Intermediate or temperature-sensitive effects are represented by a diagonal stripe. Thus, if all of the replacements destroy repressor function, then the whole bar is filled in (see position 22 or 38). If none of the replacements creates the I^- phenotype, then the whole bar is left open (see position 153 or 180). Above the 1st set of bars the replacements which create I^S repressors are shown. Here the bars were draewn only for those replacements causing this change in phenotype, the height of the bar indicating the number of replacements that generate the I^S repressor. The black segments are strong effects, and the open or dotted segments are weaker effects. The I^S phenotype can sometimes result from tighter binding to operator, as occurs for certain replacements in the amino-terminal portion (residues 1 to 61) of the repressor. These repressors are depicted as I^{tb} (tight-binding) and are represented by circles above the main set of bars. Filled-in circles indicate strong effects, and open circles weaker effects.

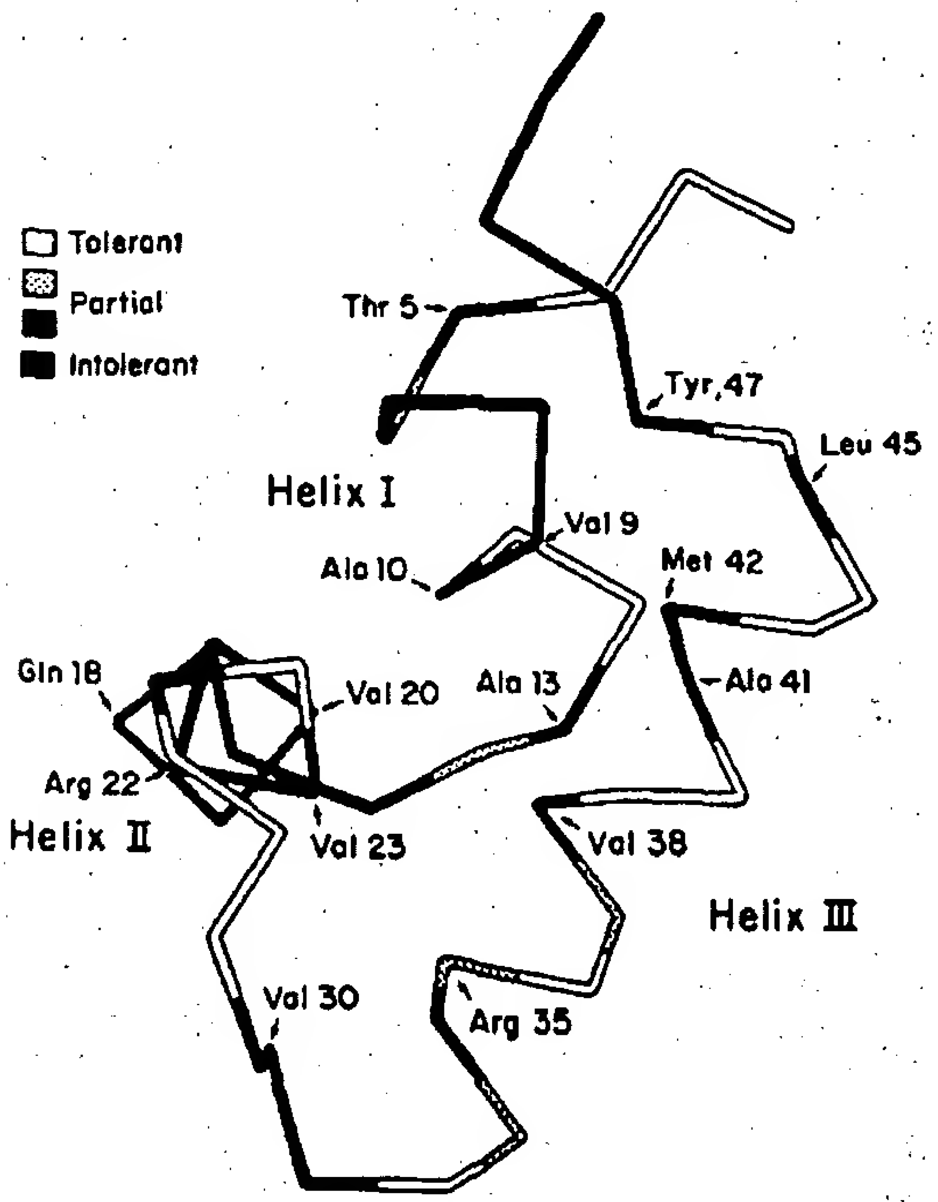


Figure 2. Three-dimensional configuration of *lac* repressor headpiece. Kaptein and co-workers have used NMR spectroscopy to determine the structure of the amino-terminal 51 amino acids of *lac* repressor (Kaptein *et al.*, 1985). Based on coordinates kindly provided by Dr R. Kaptein, we have used a Silicon Graphics Personal Iris workstation to generate this view of the repressor headpiece. The first 2 helices (I and II) are part of the helix-turn-helix motif seen in a number of regulatory proteins. The degree of tolerance to substitution, as derived from the data in Figure 1, is shown here. The residues most sensitive to substitution are shown in the darkest shading.

(c) *Homologies with related proteins*

One interesting application of the substitution data for *lac* repressor is to compare it to the observed evolutionary variation of equivalent sites of related proteins. To this end, we generated a multiple alignment of proteins homologous to *lac* repressor. FASTA (Lipman & Pearson, 1985) was used in the initial search against the PIR and translated GenBank (GenPept) protein data bases. FASTA recovered 16 proteins, all prokaryotic regulatory proteins except for the periplasmic d-ribose binding proteins of *E. coli* and *Salmonella typhimurium* (see Figures 3–4). Additional alignments of the 330 to 360 regions with eukaryotic proteins such as with *gag* polymerase and filamentous proteins were rejected after examining the pairwise alignments. Some of the matched proteins (*gal*, *cyt* repressor) have been aligned previously to *lac* repressor (von Wilcken-Bergman & Mueller-Hill, 1982). Additional searches using BLAZE and BLAST (Altschul & Lipman, 1990; Brutlag, D. L., Dautricourt, J.-P., Diaz, R., Fier, J. & Stamm, R., unpublished results) recovered 14 of the 16 proteins as highest scoring. It should be noted that other helix-turn-helix repressors such as lambda and cro do not appear in the matches. Though most of the matching proteins are bacterial repressors, only one of them, the *Klebsiella pneumoniae lac* repressor is identical in function to *E. coli lac* repressor.

Pairwise alignments between *lac* repressor and the database proteins revealed a number of conserved regions in the sequence, two of which were the most prominent. One, centered on the amino-terminal region of the proteins, comprises the helix-turn-helix DNA recognition motif. This match was found for all the proteins in the group which bind DNA. The *E. coli* periplasmic d-ribose binding protein (Groarke *et al.*, 1983) lacks this region, while the *Vibrio ScrR* protein (Blatch & Woods, 1991) consists only of this region. The second large region with extensive homologies begins at approximately residue 239 (*E. coli lac* repressor numbering) and extends at least to residue 289. It comprises a new conserved region for this family of proteins. Based on these data, a multiple sequence alignment was generated using either CLUSTALV (Higgins & Sharp, 1988) or the progressive similarity alignment of Feng & Doolittle (1990). The two programs gave very similar results. Various combinations of proteins and alignment parameters were used to ensure the robustness of the alignments. The maltose repressor (Reidl *et al.*, 1989) despite low homology in the 200 to 300 region, was aligned throughout its entire length, but the comparable region of the ribitol repressor (Wu *et al.*, 1985) could not be satisfactorily aligned. Therefore, only the first 60 amino acids of this protein was used in the alignment. No manual editing of the alignment was performed beyond these deletions. Figures 3 and 4 present the results of the alignment. Above the alignment we have placed a box whose shading indicates substitutional tolerance for that site in *E. coli lac* repressor, with intolerant residues black and tolerant ones white. This one-dimensional representation allows us to compare the conserved sites with the substitution data for *E. coli lac* repressor from Figure 1. An examination of Figures 2–3 shows a highly significant correlation between conserved sites in the alignment and sites which are intolerant to substitution in *E. coli lac* repressor, and conversely, between sites which are not conserved in evolution and sites which are tolerant to substitution. (We consider the implications of this correlation below.) For instance, there are 134 (41%) intolerant sites and 194 (59%) tolerant sites. Likewise, there are 128 (39%) conserved sites, and 200 (61%) non-conserved sites. If there were a random distribution of conserved and non-conserved sites, then among the intolerant sites we would expect 52 conserved sites, whereas we find 97, and we would expect 82 non-conserved sites, whereas we find 37. Similarly, among the tolerant residues, we would expect 76 conserved, whereas we find 31, and 118 non-conserved, whereas we find 163. These results are highly significant ($X^2 = 156$; $p \ll 0.005$). Also, the correlation coeffi-

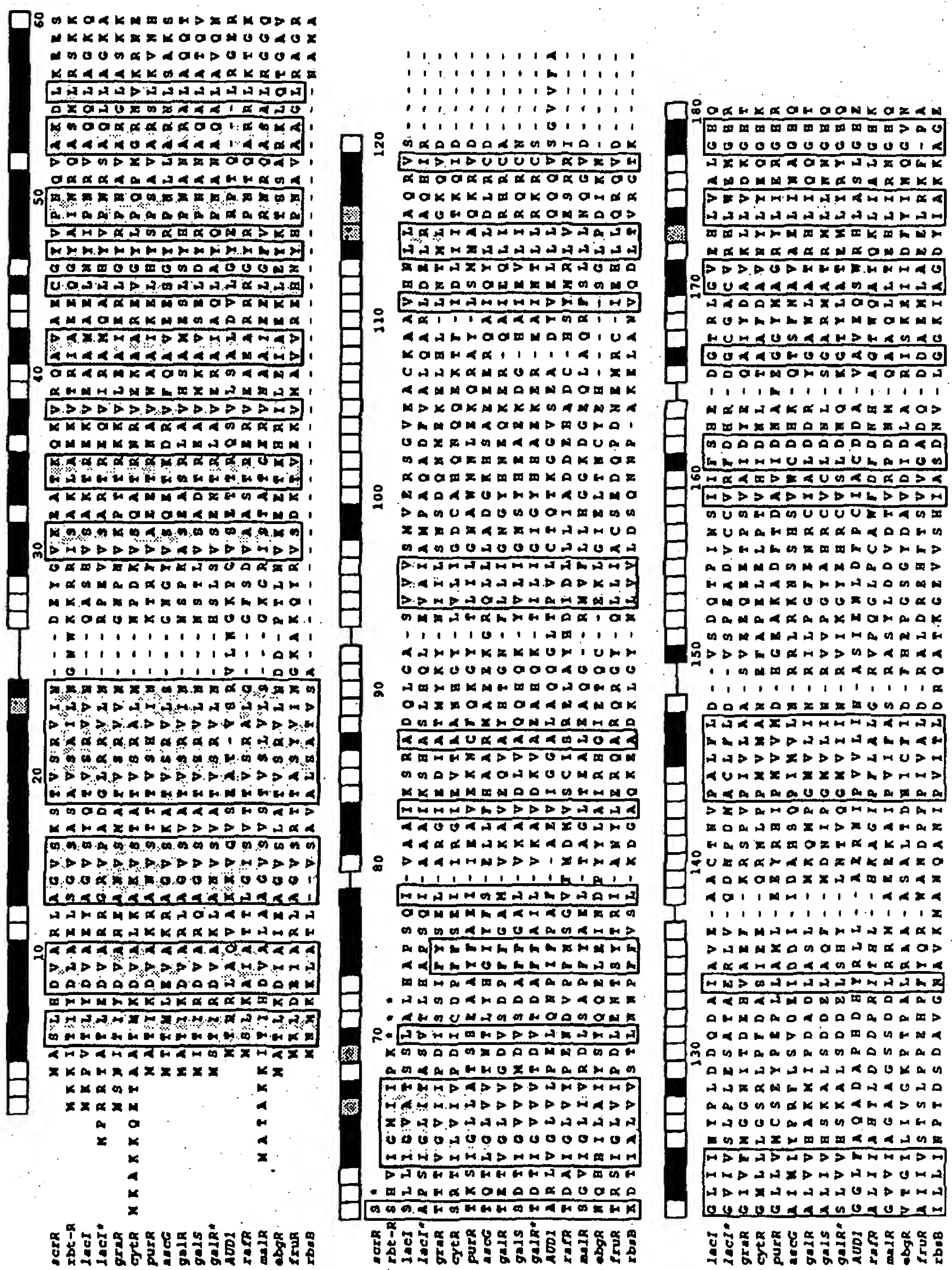


Fig. 3.

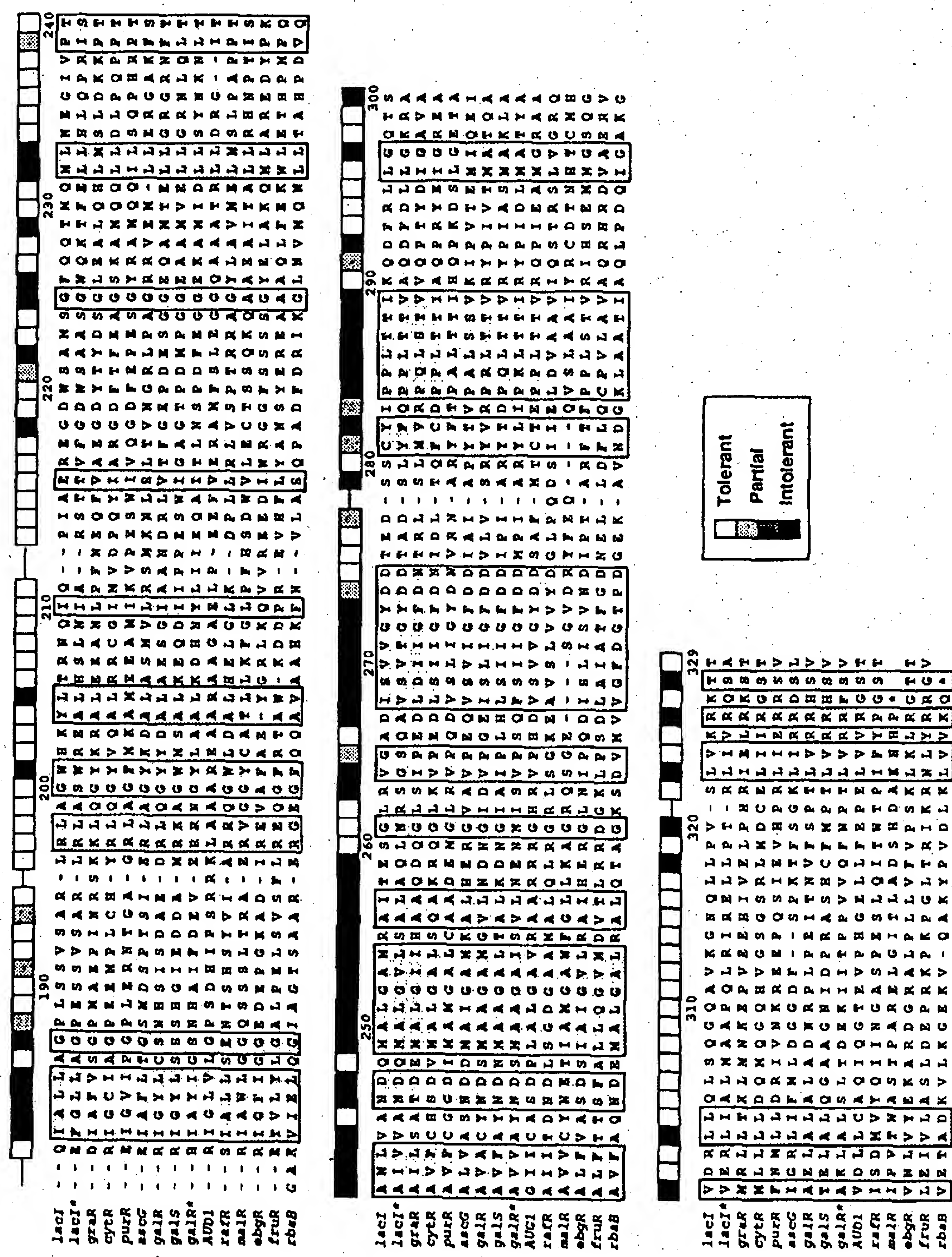


Fig. 4.

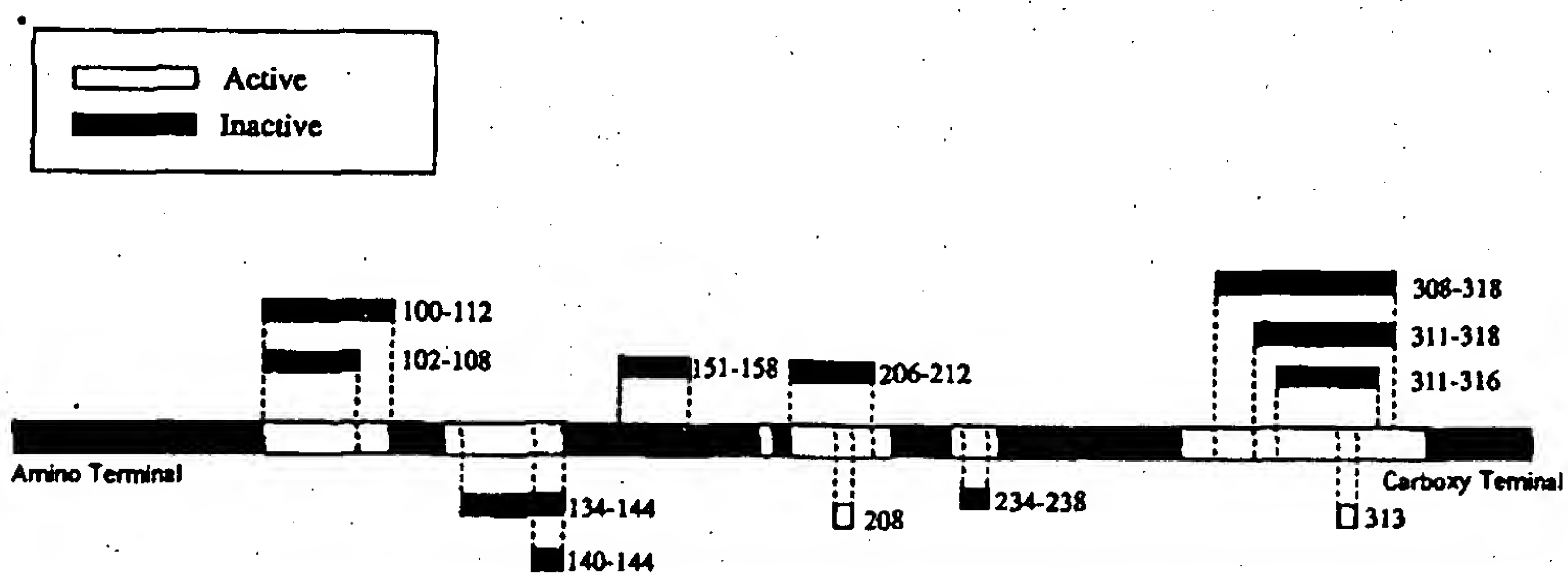


Figure 5. Deletions in the *lacI* gene. Different lengths of the *lacI* gene were deleted by site-directed mutagenesis (see Materials and Methods). The amino acids missing from the resulting repressors are shown here. Thus, delta 100–112 means that amino acids 100 to 112 are missing from the repressor. Dark bars indicate that the repressor has no activity, and an open bar (as in the case of the single amino acid deletions at positions 208 and 313) indicates that the repressor has activity (see also Table 3). The information in Figure 1 has been summarized on the main line here to indicate the degree of tolerance to substitutions of different regions of repressor. The darker the shading, the more intolerant to substitution the region is.

cient C_a (Schulz & Schirmer, 1979) is 0·58 (see Materials and Methods), which is highly significant (Matthews, 1975).

(d) *Regions of repressor that are tolerant to substitutions*

The substitution profile shows that segments of the protein are virtually insensitive to single amino replacements. What function do these stretches of up to 14 amino acids serve? To test whether these stretches could be eliminated, we deleted the residues comprising the substitution-tolerant regions using oligonucleotide-directed site-specific mutagenesis. Figure 5 shows the regions we deleted. In each case, the resulting repressor lost function *in vivo*. Namely, it could no longer repress the *lac* operon. We also constructed mutants in which we replaced the amino acid stretches with runs of alanines, as depicted in Figure 6. We used alanine because it was the most common functional replacement in our data set. In several cases replacements of up to eight amino acids with alanine runs still resulted in repressors with apparently normal function, since the *lac* genes were still IPTG inducible (see Table 3). Because some stretches can be replaced with a completely different amino acid sequence, but not deleted, it may be that these stretches serve to correctly space crucial residues in the protein, and that a unique sequence is not necessary for this function. Comparison to the three-dimensional structure, when it becomes avail-

Figures 3–4. Multiple sequence alignment of proteins with sequence homology to *lac* repressor. *scrR*, *Vibrio alginolyticus* sucrose uptake repressor (Blatch & Woods, 1991); *rbt-R*, ribitol operon repressor from *Klebsiella* (Wu *et al.*, 1985); *lacI*, *E. coli lac* repressor (Farabaugh, 1978); *lacI** (*klbR*), *Klebsiella pneumoniae lac* repressor (Buvinger & Riley, 1985); *cytR*, *E. coli cyt* repressor (Valentin-Hansen *et al.*, 1986); *purR*, *E. coli* purine biosynthesis operon repressor (Rolfes & Zalkin, 1988); *galR*, *E. coli* galactose repressor (von Wilcken-Bergman *et al.*, 1982); *ebgR*, evolved beta-galactosidase operon repressor (Stokes & Hall, 1985); *rafR*, *E. coli* raffinose operon repressor (Aslandis & Schmitt, 1990); *fruR*, fructose phosphotransferase system repressor (Jahreis *et al.*, 1991; unpublished GenBank submission X55457); *malR*, *E. coli* maltose repressor (Altshul *et al.*, 1990); *rbsB*, D-ribose periplasmic binding protein (Groarke *et al.*, 1983); *graR*, catabolite repression protein for alpha-amylase gene expression in *B. Subtilis* (Henkin, T. M., Grundy, F. J., Nicholson, W. L. & Champliss, G. H., unpublished GenBank submission M85182); *ascG*, *asc* operon regulatory protein in *E. coli* (Hall & Xu, 1992); *galS*, isorepressor of the *gal* regulon in *E. coli* (Weickert & Adhya, 1992); *galR**, *H. influenzae gal* repressor (Maskell, D. J., Szabo, M. J., Deadman, M. & Moxon, E. R., unpublished GenBank submission X65934); *AUD1*, amplification element of *S. lividans* (Piendl, W., Eichenseer, C., Viel, P., Altenbuchner, J. & Cullum, J., unpublished GenBank submission X65465). The alignments were generated using the algorithm of Feng & Doolittle (1990). Only the first 72 amino acids of the ribitol repressor were used in the alignment, since the program and CLUSTAL V (Higgins & Sharp, 1988) generated excessive gaps in the remainder of the sequence during several alignment runs. Regions of homology were boxed using a specific algorithm we developed to avoid "curve-fitting". First, for each column of amino acids at a particular site, scores from a normalized log-odds matrix (as in Feng & Doolittle, 1990) were averaged. Resulting scores ranged between 0·1 to 0·9, and columns with scores between 0·1 to 0·4 were boxed. Second, the number of the most common amino acid at any particular site was derived. The second score was used to correct for columns with log-odds scores > 0·40 that consisted of many identical amino acids (such as lysine) which have very low conservation scores in the log-odds matrix. Columns with log-odds scores between 0·4 and 0·5 for which the number of the most common amino acids was > / = 5 were boxed. Columns with log-odds scores between 0·3 and 0·4 for which the number of the most common amino acid was < 5 were unboxed. Correlation coefficients were taken from the resulting conservations and substitutional tolerance.

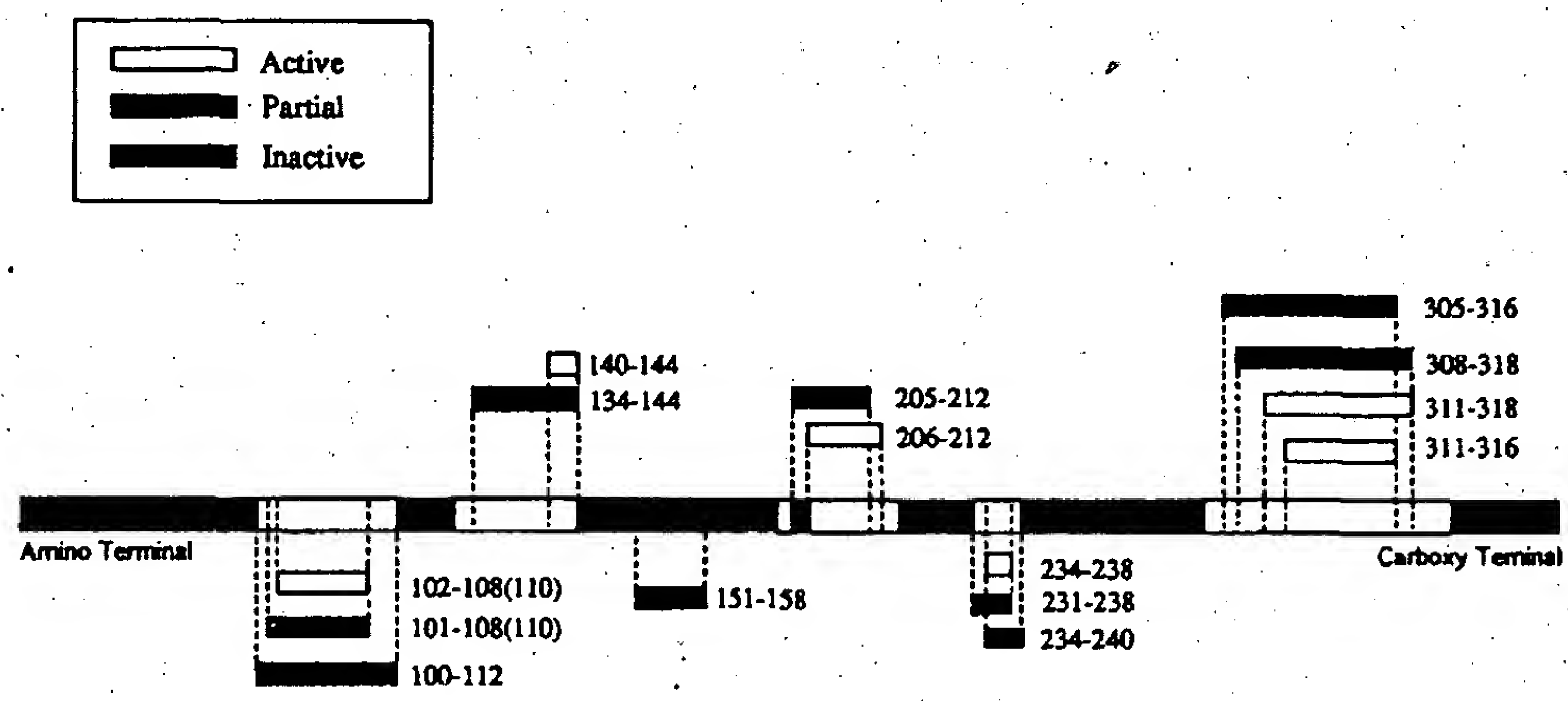


Figure 6. Alanine replacements in the *lac* repressor. Site-directed mutagenesis was used to replace segments of the *lac* repressor with continuous runs of alanines. The numbers indicate the extent of the replacements. Dark bars indicate inactive repressor, open bars indicate active repressors, and shaded bars indicate partially active repressors. See also Table 3 and the legend to Figure 5.

able, will allow the determination of whether the tolerant stretches are contained in specific structural elements, such as surface loops or helices.

(e) *Substitution patterns*

Several substitution patterns are evident. (A complete tabular representation of the substitution data will be published elsewhere). It is clear that *proline is not tolerated at many otherwise tolerant* positions. At many of the sites shown in Figure 1, only a single substitution partially or fully destroys protein function. In all cases, these represent the replacement of the wild-type residue by proline. Out of the 328 sites examined, 144 (44%) are extremely tolerant to substitutions, in that (excluding proline for the moment) they accept all 12 of the amino acids inserted by the nonsense suppressors. However, 51 (34%) of these otherwise tolerant sites *do not tolerate proline.*

Another set of sites tolerates only hydrophobic amino acids, as shown by the examples depicted in Table 4. At a number of sites, only certain small amino acids (glycine, alanine, serine and cysteine, and sometimes threonine and valine) are tolerated, as summarized in Table 5.

Table 3
Assays of β-galactosidase in strains with altered repressors

	Beta-galactosidase activity	
lacI region on plasmid	No IPTG	IPTG (10^{-3} M)
None	5600	5500
Wild-type	4·1	840
Multiple alanine replacements		
102–108	5·9	660
101–108	22	270
140–144	3·3	2700
151–158	180	850
206–212	2·4	770
234–238	4·5	1500
311–316	1·6	760
311–318	1·4	700
305–316	810	1700
Deletions		
100–112	5000	5500
206–212	8000	8000
311–318	2500	3160

Plasmids carrying each of the mutated *lacI* genes were constructed as previously described (Kleina & Miller, 1990), and put into a strain deleted for *lac* and *pro*, but carrying the *lac* region on an F′ *lacpro* episome. The episomal *lac* region carries a frameshift mutation, *1378* (Calos & Miller, 1981), in the *lacI* gene. Therefore, beta-galactosidase is synthesized constitutively from the episomal *lac* promoter, unless it is repressed by the *I* gene product synthesized from the plasmid. Beta-galactosidase was assayed after growth at 37°C. Assay conditions and units are as previously described (Miller, 1972).

(f) *Reliability of data*

The efficiency of suppression is rarely 100%, and the efficiency of suppression varies from suppressor to suppressor and from codon to codon (Miller & Albertini, 1983; Bossi, 1983), However, by using the *lacI*^{Q} allele, which results in a tenfold over-expression of the repressor (Mueller-Hill *et al.*, 1968), we can compensate for the lack of complete suppression in almost all cases. For the majority of suppressed proteins, the amount of repressor being produced varies within a threefold range, near to the normal amounts of wild-type repressor, since we are using a single copy F′ carrying the *lacI* gene. Thus, we are not creating false positives by greatly overproducing the repressor. Also, the assays we are using (Kleina & Miller, 1990) to determine repressor function *in vivo* can recognize, as partially defective, repressors with as much as 8 to 10% activity (see Materials and Method).

Leakiness of the amber mutation is not a factor in the experiments reported here. All amber mutations allow some level of transmission even in strains lacking known suppressors. Studies of fusion strains estimate this level to vary between 0·01% and 2%

Table 4
Substitution patterns of selected sites in the repressor

Wild-type	Amino acid appearing at position: Ile	Leu	Phe	Val	Pro	Cys	Tyr	Ala	Gly	Ser	His	Gln	Arg	Lys	Glu
Residue															
Ile64	(+)	+	+		+	+	+	+	−	−	+	−	−	−	−
Val66		+	+, s	(+)	−	+	+, s	+	−	−, ts	+	−	−	−	−
Leu71		(+)	+			−	+	−	−	−		−, ts	−	−	−
Ile123	(+)	+	+	+	−	+	−	+	−	−	−	−	−	−	−
Ile124	(+)	+	+	+	−	+	+, ts	±	−	−	+	−	−	−	−
Phe147		+	(+)		−	±	+	+	−	−	+	−, ts	−	−	−
Phe161		+, ts	(+)		−, ts	−		−	−	−, ts	+	−	−	−	−
Leu174		(+)	+		−	+	+	+	−	−, ts	−	−, ts	−		−
Ile182	(+)	+	+	+	+	+	+	+	−	−	−	−	−	−	−
Leu243		(+)	+		−	+	−	±	−	−	−	−	−	−	−
Val270		+	+	(+)	−	+	−	+	−	−	−	−	−	−	−
Leu286		(+)	−	+	−	±	−	±, ts	−	−	−	−	−	−	−
Ile289	(+)	+		+	−	+	−	−	−	−	−	−	−	−	−

All amino acid replacements were made by using the amber suppressors indicated in Table 1. See also Kleina & Miller (1990). Different designations are used for the I phenotype. In general, + refers to no significant detected alteration (greater than 200-fold repression of beta-galactosidase in cases where measured); ± indicates altered repression, but retention of the ability to repress beta-galactosidase synthesis 20 to 200-fold; − usually designates less than 4-fold repression. These designations are only a rough guide. Substitutions resulting in a temperature-sensitive phenotype are indicated by ts. Substitutions resulting in a loss of response to inducer are indicated as s for I^s repressors, or ws for a weaker I^s phenotype.

(Miller & Albertini, 1983). However, even with the overproducing I^Q allele operating, none of the amber mutations in the *lacI* gene (from codons 2 to 329) exhibit any measurable repressor activity in the absence of a suppressor under any experimental conditions we have employed.

(g) *Implications*

The work of Perutz and co-workers on the hemoglobins and myoglobins (Perutz *et al.*, 1965; Perutz & Lehmann, 1968) established that in globular proteins the non-polar residues in the interior of the protein are not replaceable with polar amnino acids, but in many cases are replaceable with certain other non-polar amino acids. On the other hand, residues on the surface of the protein are usually freely exchangeable between non-polar and polar amino acids, unless they are part of the substrate or ligand binding site, or part of intersubunit contacts. Among similar proteins in different species, residues at interior positions are more highly conserved than residues on the surface (Pertuz *et al.*, 1965). Work with extensive amino acid replacements in other proteins has reinforced these conclusions. Sauer and co-workers (Bowie *et al.*, 1990) have found, using combinatorial cassettee mutagenesis, that many residues in the phage lambda repressor amino-terminal domain can be freely substituted, but buried residues are more refractory to amino acid

Table 5
Effects of amino acid replacements at selected sites in the repressor

Wild-type	Amino acid appearing at position: Leu	Phe	Pro	Cys	Tyr	Ala	Gly	Ser	Thr	His	Gln	Arg	Lys	Glu
Residue														
Ala57	−	−	−		−	(+)	±			−	−	−	−	−
Gly65	−	−	−	−	−	−	(+)	−		−	−	−	−	−
Gly166	−	−	−	±, ts	−	+	(+)	+		−	−	−	−	−
Gly218	±, ts	−	−	+	−	+	(+)	+		−	−	−	−	−
Gly225	−	−	−	−	−	+	(+)	±, ts		−	−	−	−	−
Ala241	−, ts	−, ts	−	+	−	(+)	+	+		−	−	−	−	−
Ala250	−	−	−	+	−	(+)	+	+		−	−	−	−	−
Gly252	−	−	−	−	−	−	(+)	+		−	−	−	−	−
Gly272	−	−	−	−	−	+	(+)	−		−	−	−	−	−
Ser279	−	−	−	+, ws	−	+, ws	+, s	(+)		−	−	−	−	−
Thr287	−	−	−	+	−	+	−	+	(+)	−	−	−	−	−
Thr288	−	−	−	+	−	+	+	+	(+)	−	−	−	−	−
Gly297	−, ws	−	−	−	−, ws	±	(+)	±		−	−	−	−	−
Thr328	−	−	−	−	−	±	+	+	(+)	−	−	−	−	−

For details, see legend to Table 4.

exchanges. Only positions on the surface could tolerate replacements by polar residues. A recent study of the systematic replacement of amino acids at 163 of the 164 positions in bacteriophage T4 lysozyme (Rennell *et al.*, 1991), using our amber suppressor system, has shown a very strong correlation between degree of substitutional tolerance and the degree of solvent accessibility. Residues which are exposed to solvent (and on the surface) are freely substitutable, whereas those which are not accessible (and buried) are intolerant to substitution. In light of these and our previous studies (Miller *et al.*, 1979; Kleina & Miller, 1990), the finding that a large number of sites in the repressor can be freely substituted is not surprising. In the repressor work reported here, 28% of the 328 sites tolerated all substitutions, 44% all substitutions except proline, and 59% most substitutions. The T4 lysozyme study revealed that 55% of the sites tolerated all of the substitutions. The somewhat higher percentage of sites tolerating every substitution can be partly attributed to the different sensitivities of the assays used (Rennell *et al.*, 1991; see Materials and Methods). The form of the results in the two data sets is strikingly similar. For example, in both the repressor and lysozyme study, proline is frequently not tolerated at otherwise tolerant sites. Extensive substitutions have also been used to identify important functional regions in other proteins. For examples, random mutagenesis and screening was used to detect missense mutations in each of the 99 coding positions in the HIV-1 protease, which revealed three important functional regions of the protein (Loeb *et al.*, 1989).

One striking feature of Figure 1 is how it reveals the regions which are sentitive to substitution. Two sizeable regions, from residues 1 to 59 and from residues 241 to 289, are very sensitive to substitutions, creating the I^- phenotype. The amino-terminal 59 amino acids form a separate domain which includes the DNA and operator binding sites (Mueller-Hill, 1975; Platt *et al.*, 1973; Ogata & Gilbert, 1978), which explains its intolerance to many amino acid substitutions. The second, 241 to 289 region, has an unknown function. It is precisely this segment which shown the strongest conservation among proteins related to the repressor (Figures 2–3). Note that this region also includes several clusters of I^S sites, presumably defining part of the inducer binding site.

A second striking feature of Figure 1 is the presence of segments which are almost completely tolerant of single amino acid substitutions. This includes residues 100 to 112, 129 to 145, 151 to 160, 206 to 217 and 305 to 318. These "open" regions are interdispersed between single residues or clusters of residues that are sensitive to substitution. The finding that the residues in the open regions cannot be deleted (Figure 5) but can be replaced by poly-alanine stretches (Figure 6) indicates that stretches of the sequence serve as spacers for the hydrophobic residues in the core. In fact, in the vast majority of cases the sensitive residues separated by the spacers are hydrophobic amino acids, and the suggestion is that they form part of the hydrophobic core of the protein. Although the three-dimensional structure of the *lac* repressor is not yet known, X-ray crystallographic data on the structure is forthcoming (Pace *et al.*, 1990). The data in Figures 1, 5 and 6 predict that the open regions which are replaceable by poly-alanine stretches may represent surface residues, since they can accept replacements of up to eight amino acids at a time. It should be noted that a method called "clustered charged-charged-to-alanine scanning mutagenesis" (Bennett *et al.*, 1991; Bass *et al.*, 1991, see also Wertman *et al.*, 1992) surveys the surface of the protein by substituting alanines for all the charged residues within a five amino acid stretch. Our replacements complement the recent study of Matthews and co-workers (Heinz *et al.*, 1992), who replaced ten consecutive alanine residues in bacteriophage T4 lysozyme, in a segment known to form an alpha-helix on the exterior of the protein. The lysozyme retained normal function and stability. Our work differs in making the replacements using functional, rather than structural, criteria, and may therefore demonstrate other types of structure tolerant to substitutions.

It is interesting to examine the nature of substitutionally intolerant residues in the repressor. The amino-terminal 59 amino acids are particularly sensitive to amino acid replacements. However, if we look at the remainder of the repressor (actually from residues 60 to 329), it is evident that virtually all sensitive sites are either hydrophobic residues that can only be replaced by certain other hydrophobic residues, or else small amino acids that can only be substituted by other small amino acids. Table 6 shows that only 6 of 73 hydrophilic residues in this region are sensitive to substitution, whereas 42 of 97 hydrophobic residues and 28 of 93 small residues are sensitive to substitution.

There is a strong, although not complete, correlation (see Materials and Methods) between the evolutionarily conserved regions of the repressor and the residues that are sensitive to substitution, as can be seen from Figures 2–3. In others words, the substitution map of the protein is strongly correlated with the "evolution map" of the *lac* repressor family. Considering that all but one of the proteins being compared are not *lac* repressors but other regulatory proteins, this correlation is remarkable. One possible implication of this correlation is that the sequence divergence of these proteins resulted from near neutral mutations which occurred relatively independently, with multiple compensating mutations being rare. In other words, the pattern of allowed amino acids for a particular site in a protein may remain constant even when up to 80% of the amino acids in the protein have changed during evolutionary divergence. A similar study of suppressor-generated changes in phage T4 lysozyme showed that whereas 74 of 163 residues tested (45%) are sensitive to at least one substitution, all 14 residues that are fully conserved among five phage-encoded lysozymes are sensitive to

substitutions (Poteete *et al.*, 1992). It should be noted that although the variability of a site generally correlates with the evolutionary variability, this is not always the case. For instance, Tyr17 and Gln18 in the amino-terminal region of the repressor are intolerant to substitution yet vary considerably in the aligned proteins. These sites have both been implicated in controlling the specificity of DNA binding (Lehming *et al.*, 1987, 1988). Since operator sequences vary among DNA binding proteins, one expects similar variation of amino acids conferring specificity. With respect to the above, it is interesting to note that the I^S sites (residues presumably involved in inducer binding) are not well conserved. One might rationalize this result by arguing that the specific ligands being bound by each regulatory protein are different, so that parts of the binding site involved in specificity should be variable in the *lac* repressor family, but intolerant to substitutions in any given protein from the set, for the ligand binding property.

There are also a number of residues, such as Gly91, which are highly conserved in the aligned set but can be freely substituted in the repressor. In some cases these residues represent sites which simply do not conform to the general pattern, and in other cases may be due to specious homologies or unrecognized subtle functions of the repressor protein.

We thank Drs Ponzy Lu, Mitch Lewis, David Eisenberg, and Brian Matthews for valuable discussions. We are indebted to Dr Robert Kaptein for supplying the coordinates of the *lac* repressor headpiece shown in Figure 2. This work was supported by grants from the National Institute of Health (USHHS RO1 GM 43827-02) and from the Department of Defense (DAALO3-92-G-0173).

References

Alberti, S., Oehler, S., Wilcken-Bergmann, B. v., Kraemer, H. & Mueller-Hill, B. (1991). Dimer-to-tetramer assembly of *lac* repressor involves a leucine heptad repeat. *New Biol.* 3, 57–62.

Altshul, S. F. & Lipman, D. J. (1990). Protein database searches for multiple alignments. *Proc. Nat. Acad. Sci., U.S.A.* **87**, 5509–5513.

Aslanidis, C. & Schmitt, R. (1990). Regulatory elements of the raffinose operon: nucleotide sequences of operator and repressor genes. *J. Bacteriol.* **172**, 2178–2180.

Bass, S. H., Mulkerrin, M. G. & Wells, J. A. (1991). A systematic mutational analysis of hormone-binding determinants in the human growth hormone receptor. *Proc. Nat. Acad. Sci., U.S.A.* **88**, 4498–4502.

Bennett, W. F., Paoni, N. F., Keyt, B. A., Botstein, D., Jones, A. J., Presta, L., Wurm, F. M. & Zoller, M. J. (1991). High resolution analysis of functional determinants on human tissue-type plasminogen activator. *J. Biol. Chem.* **266**, 5191–5201.

Blatch, G. L. & Woods, D. R. (1991). Nucleotide sequence and analysis of the *Vibrio alginolyticus scr* repressor-encoding gene (*scrR*). *Gene*, **101**, 45–50.

Bossi, L. (1983). Context effects: translation of UAG codon by suppressor tRNA is affected by the sequence following UAG in the message. *J. Mol. Biol.* **164**, 73–87.

Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A. & Sauer, R. T. (1990). Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science*, **247**, 1306–1310.

Bradley, D., Park, J. V. & Soll, L. (1981). $tRNA_2^{Gln}Su^{+2}$ mutantrs that increase amber suppression. *J. Bacteriol.* **145**, 704–712.

Buvinger, W. E. & Riley, M. (1985). Regulatory region of the divergent *Klebsiella pneumoniae lac* operon. *J. Bacteriol.* **163**, 858–862.

Calos, M. P. & Miller, J. H. (1981). Genetic and sequence analysis of frameshift mutations induced by ICR-191. *J. Mol. Biol.* **153**, 39–66.

Chakerian, A. E., Tesmer, V. M. Manly, S. P., Brackett, J. K., Lynch, M. J., Hoh, J. T. & Matthews, K. S. (1991). Evidence for Leucine zipper motif in lactose repressor protein. *J. Biol. Chem.* **266**, 1371–1374.

Chamness, G. C. & Willson, C. D. (1970). An unusual *lac* repressor mutant. *J. Mol. Biol.* **53**, 561–565.

Farabaugh, P. J. (1978). Sequence of the *LacI* gene. *Nature (London)*, **274**, 765–769.

Feng, D. F. & Doolittle, R. F. (1990). Progressive alignment and phylogenetic tree construction of protein sequences. *Methods Enzymol.* **183**, 375–387.

Groarke, J. M., Mahoney, W. C., Hope, J. N., Furlong, C. E., Robb, F. T., Zalkin, H. & Hermodson, M. A. (1983). The amino acid sequence of D-ribose-binding protein from *Escherichia coli* K12. *J. Biol. Chem.* **258**, 12952–12956.

Hall, B. G. & Xu, L. (1992). Nucleotide sequence, function, activation, and evolution of the cryptic *asc* operon of *Escherichia coli* K12. *Mol. Biol. Evol.* **9**, 688–706.

Heinz, D. W., Baase, W. A. & Matthews, B. W. (1992). Folding and function of a T4 lysozyme containing 10 consecutive alanines illustrate the redundancy of information in an amino acid sequence. *Proc. Nat. Acad. Sci., U.S.A.* **89**, 3751–3755.

Higgins, D. G. & Sharp, P. M. (1988). CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene*, **73**, 237–244.

Jahreis, K., Postma, P. W. & Lengeler, J. W. (1991). Nucleotide sequence of the ilv H-fruR gene of *Escherichia coli* K-1 and *Salmonella typhimurium* LT2. *Mol. Gen. Genet.* **226**(1-2), 332–336.

Kaptein, R., Zuiderweg, E. R. P., Scheek, R. M., Boelens, R. & Van Gunsteren, W. F. (1985). A protein structure from nuclear magnetic resonance data. *J. Mol. Biol.* **182**, 179–182.

Kleina, L. G. & Miller, J. H. (1990). Genetic studies of the *lac* repressor. Extensive amino acid replacements generated by the use of natural and synthetic nonsense suppressors. *J. Mol. Biol.* **212**, 295–318.

Kleina, L. G., Masson, J.-M., Normanly, J., Miller, J. H. & Abelson, J. (1990). Construction of *Escherichia coli* amber suppressor tRNA Genes. II. Synthesis of additional tRNA genes and improvement of suppressor efficiency. *J. Mol. Biol.* **213**, 705–717.

Lehming, N., Sartorius, J., Oehler, S., Wilcken-Bergman, B. v. & Mueller-Hill, B. (1987). The interaction of the recogition helix of *lac* repressor with *lac* operator. *EMBO J.* **6**, 3145–3153.

Lehming, N., Sartorius, J., Oehler, S., Wilcken-Bergman, B. v. & Mueller-Hill, B. (1988). Recognition helices of *lac* and lambda repressor are oriented in opposite

directions and recognize similar DNA sequences. *Proc. Nat. Acad. Sci., U.S.A.* **85**, 7947–7951.

Lipman, D. J. & Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science*, **227**, 1435–1441.

Loeb, D. D., Swanstrom, R., Everitt, L., Manchester, M., Stamper, S. E. & Hutchinson, C. A., III (1989). Complete mutagenesis of the HIV-1 protease. *Nature (London)*, **340**, 397–400.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. (The Netherlands)*, **405**, 442–451.

McClain, W. H. & Foss, K. (1988). Changing the acceptor identity of a transfer RNA by altering nucleotides in a variable pocket. *Science*, **241**, 1804–1807.

Mckay, D. B. & Steitz, T. A. (1981). Structure of catabolite gene activator protein at 2·9 Å resolution suggests binding to left-handed *B*-DNA. *Nature (London)*, **290**, 744–749.

Miller, J. H. (1992). *Experiments in Molecular Genetics*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Miller, J. H. (1978). A review. *The Operon* (Miller, J. H. & Reznikoff, W. S., eds), pp. 31–88, Cold spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Miller, J. H. & Albertini, A. M. (1983). Effects of surrounding sequence on the suppression of nonsense codons. *J. Mol. Biol.* **164**, 59–71.

Miller, J. H. & Schmeissner, U. (1979). Analysis of missense mutations in the *lacI* gene. *J. Mol. Biol.* **131**, 223–248.

Miller, J. H., Coulondre, C., Hofer, M., Schmeissner, U., Sommer, H., Schmitz, A. & Lu, P. (1979). Generation of altered proteins by the suppression of nonsense mutations. *J. Mol. Biol.* **131**, 191–222.

Mueller-Hill, B. (1975). A review. *lac* repressor and *lac* operator. *Prog. Biophys. Mol. Biol.* **30**, 227–252.

Mueller-Hill, B., Crapo, L. & Gilbert, W. (1968). Mutants that make more *lac* repressor. *Proc. Nat. Acad. Sci., U.S.A.* **59**, 1259–1264.

Myers, G. L. & Sadler, J. R. (1971). Mutational inversion of control of the lactose operon of *E. coli*. *J. Mol. Biol.* **58**, 1–28.

Normanly, J., Kleina, L. G., Masson, J.-M., Abelson, J. & Miller, J. H. (1990). Construction of *Escherichia coli* amber suppressor tRNA *genes*. III. Determination of tRNA specificity. *J. Mol. Biol.* **213**, 719–726.

Normanly, J., Masson, J.-M., Kleina, L. G., Abelson, J. & Miller, J. H. (1986). Construction of two *Escherichia coli* amber suppressor genes: tRNA Phe/Cua and tRNA Cys/Cua. *Proc. Nat. Acad. Sci., U.S.A.* **83**, 6548–6552.

Ogata, R. & Gilbert, W. (1978). An amino-terminal fragment of *lac* repressor binds specifically to *lac* operator. *Proc. Nat. Acad. Sci., U.S.A.* **75**, 5851–5854.

Pace, H. C., Lu, P. & Lewis, M. (1990). *lac* repressor: crystallization of intact tetramer and its complexes with inducer and operator DNA. *Proc. Nat. Acad. Sci., U.S.A.* **87**, 1870–1873.

Perutz, M. F. (1965). Structure and function of haemoglobin. I. tentative atomic model of horse oxyhaemoglovin. *J. Mol. Biol.* **13**, 646–668.

Perutz, M. F. & Lehmann, H. (1968). Molecular pathology of human haemoglobin. *Nature (London)*, **219**, 902–909.

Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965). Structure and function of haemoglobin. II. Some relations between polypeptide chain configuration and amino acid sequence. *J. Mol. Biol.* **13**, 669–678.

Platt, T., Files, J. G. & Weber, K. (1973). Specific proteolytic destruction of the NH_2-terminal region and loss of the deoxyribonucleic acid-binding activity. *J. Biol. Chem.* **248**, 110–121.

Platt, T., Miller, J. H. & Weber, K. (1970). *In vivo* degradation of mutant *lac* repressor. *Nature (London)*, **228**, 1154–1156.

Poteete, A. R., Rennell, D. & Bouvier, S. E. (1992). Functional significance of conserved amino acid residues. *Proteins*, **13**, 38–40.

Reidl, J., Roemisch, K., Ehrmann, M. & Boos, W. (1989). Mal I, a novel protein involved in regulation of the maltose system for *Escherichia coli*, is highly homologous to thge repressor proteins GalR, CytR and *Lac*I. *J. Bacteriol.* 171, 4888–4899.

Rennell, D., Bouvier, S. E., Hardy, L. W. & Poteete, A. R. (1991). Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol.* **222**, 67–87.

Rolfes, R. J. & Zalkin, H. (1988). *Escherichia coli* gene purR encoding a repressor protein for purine nucleotide synthesis: cloning, nucleotide sequence, and interaction with the purF operator. *J. Biol. Chem.* **263**, 19653–19661.

Schmitz, A., Schmeissner, U., Miller, J. H., & Lu, P. (1976). Mutations affecting the quaternary structure of the *lac* repressor. *J. Biol. Chem.* **251**, 3359–3366.

Schulz, G. E. & Schirmer, R. H. (1979). *Principles of Protein Structure*, pp. 124–125, Springer-Verlag, New York.

Stokes, H. W. & Hall, B. G. (1985). Sequence of the ebgR gene of *Escherichia coli*: evidence that the EBG and LAC operons are descended from a common ancestor. *Mol. Biol. Evol.* **2**, 478–483.

Valentin-Hansen, P., Larsen, J. E. L., Hojrup, P., Short, S. A. & Barbier, C. S. (1986). Nucleotide sequence of the CytR regulatory gene of *E. coli* K-12. *Nucl. Acids Res.* **14**, 2215–2228.

Weickert, M. J. & Adhya S. (1992). Isorepressor of the gal regulon in *Escherichia coli*. *J. Mol. Biol.* **226**, 69–83.

Wertman, K. F., Drubin, D. G. & Botstein, D. (1992). Systematic mutational analysis of the yeast *ACTI* gene. *Genetics*, **132**, 337–350.

Wilcken-Bergman, B. v. & Mueller-Hill, B. (1982). Sequence of *galR* gene indicates a common evolutionary origin of *lac* and *gal* repressor in *Escherichia coli*. *Proc. Nat. Acad. Sci., U.S.A.* **79**, 2427–2431.

Wu, J., Anderson-Loviny, T., Smith, C. A. & Hartley, B. S. (1985). Structure of wild-type and mutant repressors and of the control region of the *rbt* operon of *Klebsiella aerogenes*. *EMBO J.* **4**, 1339–1344.

Edited by B. W. Matthews

(Received 30 April 1993; accepted 18 April 1994)

Biochemistry **1998**, *37*, 7157−7166

A Search for Single Substitutions That Eliminate Enzymatic Function in a Bacterial Ribonuclease†

Douglas D. Axe,* Nicholas W. Foster, and Alan R. Fersht*

Cambridge University Chemical Laboratory and Cambridge Centre for Protein Engineering, MRC Centre, Hills Road, Cambridge CB2 2QH, U.K.

Received February 19, 1998

ABSTRACT: Exhaustive-substitution studies, where many amino acid replacements are individually tested at all positions in a natural protein, have proven to be very valuable in probing the relationship between sequence and function. The broad picture that has emerged from studies of this sort is one of functional tolerance of substitution. We have applied this approach to barnase, a 110-residue bacterial ribonuclease. Because the selection system used to score barnase mutants as active or inactive detects activity down to a level that can be approached by nonenzyme catalysts, mutants that test inactive are essentially devoid of enzymatic function. Of the 109 barnase positions subjected to substitution, only 15 (14%) are vulnerable to this extreme level of inactivation, and only 2 could not be substituted without such inactivation. A total of 33 substitutions (amounting to 5% of the explored substitutions) were found to render barnase wholly inactive. The profoundly disruptive effects of all of these inactivating substitutions appear to result from either (1) replacement of a side chain that is directly involved in substrate binding or catalysis, (2) replacement of a substantially buried side chain, (3) introduction of a proline residue, or (4) replacement of a glycine residue. Although substitutions of these types are functionally tolerated more often than not, the system used here indicates that only these sorts of substitution are capable of single-handedly reducing catalytic function to, or nearly to, levels that can be achieved by nonenzyme catalysts.

It is hoped that investigations of protein folding will ultimately yield a comprehensive understanding of the relationship between protein sequence and structure. Although this is an ambitious undertaking, it is only half of the larger effort aimed at elucidating the relationship between sequence and function. The other half, aimed at understanding the structure−function relationship, is no less ambitious.

While a general solution to the grand problem of relating protein sequence to function will clearly be some time in the making, we do have at our disposal, in the form of natural proteins, thousands of specific solutions to this problem. It therefore makes good sense for us to glean as much raw data as we possibly can from these natural solutions. Given the complexities and subtleties of the sequence−function relationship, it also makes sense for us to collect these data in a manner that is entirely unbiased by any a priori expectations we may hold regarding the nature of that relationship.

The simplest and most direct way to obtain such an unbiased data set is to produce and test a large collection of mutant proteins where all possible single-residue substitutions are represented. This might be termed the exhaustive-substitution approach. Since all single-replacement possibilities are examined by this approach, the data lack any imprint of the experimenter's expectations, and they are complete in the sense that the entire molecule is examined (there is no question that more information would be gained by examining all possible double or triple substitutions, but such increases in the level of substitution lead very quickly to impractically large numbers of mutants).

A number of studies have captured the essence of this approach (*1*−*8*). One of the broad features to emerge from these studies (*1*, *4*, *5*, *7*) and others (*9*, *10*) is a positive correlation between the degree of solvent exposure at an amino acid position and the level of substitutional tolerance at that position. It has also become clear that natural proteins typically tolerate single substitutions, even nonconservative ones, at most positions without complete loss of function. A peculiar exception to this is the phage P22 Arc repressor, which was found to tolerate nonconservative substitutions at only 8 of its 53 positions (*2*). Perhaps the unusual behavior of this protein can be attributed in part to its unusually small size.

Another factor, one that clearly affects the results of any exhaustive-substitution study, is the activity threshold, the minimum level of activity necessary for a mutant to be scored active. Because of the large number of mutants involved, these studies typically rely upon rapid in vivo screens that produce a binary (i.e., "active" or "inactive") indication of activity. It is generally possible to set the activity threshold at various levels within some range, the choice being more a matter of experimental convenience than necessity. For any experimental protein, a high threshold will lead to a more inclusive list of positions deemed to be functionally important than will a low threshold. In previous studies, thresholds

† This work was supported in part by a grant from the Office of Naval Research.

* Corresponding authors. Fax: 1223 402140.

S0006-2960(98)00402-4 CCC: $15.00
Published on Web 05/02/1998

have been in the range of 3−30% of wild-type (WT)[1] activity (*1*, *2*, *4*−*6*).

Barnase, a bacterial ribonuclease, provides an opportunity to apply the exhaustive-substitution approach using a very low activity threshold. The extreme autotoxicity of this enzyme (*11*) allows direct selection of mutants with very low (approximately 0.1% of wild-type and lower) activities (*12*), thereby enabling us to directly identify residues that profoundly influence enzyme function.[2] Here we report the results of an experiment designed to identify all single-base missense mutations that affect barnase function to this extent.

MATERIALS AND METHODS

Strains and Plasmids. The *Escherichia coli* strains and the plasmid used in this work have been described previously (*12*). Plasmid pSNBR carries a synthetic barnase gene, *synbar*, that is interrupted by two amber stop codons. Strain C-1a, a nonsupressing strain, is used to prepare pSNBR DNA. Strain MX383, a suppressing strain, reads amber stop codons as serine codons, causing a full-length product to be produced.

Mutagenesis of synbar. The number of possible single-position substitutions for a protein the size of barnase is large enough to make individual preparation, sequencing, and testing of all single mutants impractical. Various approaches, all with strengths and weaknesses, have been employed in previous studies to overcome this difficulty. Three constraints were dominant in our choice of a method for mutagenesis. First, since this is a study of the effects of single substitutions, we required that the chosen method primarily produce singly substituted variants. Second, since the *synbar* selection system selects clones producing inactive barnase variants, we needed a method that would minimize accidental introduction of frame-shift mutations (virtually all of which would pass selection). And third, since a particular sort of amber suppression (encoded by the *supD* allele) is an integral part of the *synbar* selection system (*12*), methods involving a large set of strains with various amber-suppression phenotypes (*1*) could not be used.

These constraints can be adequately satisfied by a method that uses mutagenic oligonucleotides containing suitably small amounts of "contaminating" bases. By incorporating these oligonucleotides in a manner that avoids blunt-end ligations, we are able to achieve a low background frequency of frame-shift mutations. The primary limitation to this approach is the fact that it restricts the substitution set at a particular position to the amino acids that can be specified with a codon that differs only by one base from the wild-type codon. For *synbar*, this means that the number of accessible substitutions per position ranges from 5 to 7 (average = 6.2). Although most substitutions are inaccessible by this method, the number and variety of accessible substitutions ensure that multiple nonconservative substitutions (in addition to more conservative ones) will be possible at all positions. This will enable us to obtain the desired information.

The coding region of *synbar* was conceptually divided into 8 contiguous regions covering from 12 to 14 codons each, starting from the second codon.[3] With each region being treated separately, random base substitutions were introduced throughout the gene by using oligonucleotides prepared with mixed phosphoramidites. For each region, an oligonucleotide was synthesized that spanned the entire region and extended 10 bases beyond in both directions. The 10-base extensions were synthesized so as to perfectly complement the corresponding regions on the template plasmid, pSNBR. The central portion of each oligonucleotide, corresponding to one of the 8 *synbar* regions, was synthesized with small concentrations of contaminating phosphoramidites to introduce base substitutions at a low frequency. This was achieved by preparing the four standard phosphoramidite solutions at double concentrations and transferring 118 μL from each of these bottles to a fifth bottle containing 20 mL of pure anhydrous acetonitrile. The synthesizer was programmed to draw only from the pure bottles for the first and last 10 base positions of each oligonucleotide but to draw equal volumes from the appropriate pure bottle and the mixed bottle at each of the central positions. The resulting product is a mixed population of oligonucleotides where each particular mutation occurs as frequently in isolation as it does in combination with other mutations (the latter being undesirable for this work).

In separate PCRs, each of the 8 mutagenic oligonucleotides was used with a single biotinylated oligonucleotide to amplify a large portion of plasmid pSNBR (Figure 1). After purification from agarose gels, the product DNA was methylated by incubation with *E. coli* Dam methylase. The nonbiotinylated strands were then separated from the biotinylated strands by using Dynabeads M-280 Streptavidin (Dynal) according to the manufacturer's protocol. The nonbiotinylated mutant strands were retained for producing mutant plasmid clones. Two additional oligonucleotide primers, one biotinylated, were used to amplify a portion of pSNBR such that the nonbiotinylated strand from this product can anneal to any of the nonbiotinylated mutant strands to form a gapped-duplex plasmid molecule (Figure 1). After isolation of the nonbiotinylated strand as before, the gapped-duplex product was prepared by combining this strand with each mutant strand (in equal proportions), heating to 75 °C for several minutes, and allowing the mixtures to cool slowly to room temperature.

The synbar Selection System. E. coli strain MX383 was transformed directly [by the previously described protocol (*12*)] with the mixtures of gapped-duplex DNA. Dam methylation of the mutant strand causes the cell to use this strand as the reference in correcting mismatches (*14*), thereby ensuring that *synbar* mutations are preserved. Because of the extreme autotoxicity of *synbar* expression (*11*, *12*), only mutant genes producing largely inactive barnase variants

[1] Abbreviations: fMet, *N*-formylmethionine; FMOC, 9-fluorenylmethoxycarbonyl; HPLC, high-performance liquid chromatography; PCR, polymerase chain reaction; RNase, ribonuclease; WT, wild-type.

[2] To avoid confusion, it should be emphasized that throughout this work the terms inactive, inactivating, functional, sensitive, tolerant, and related words refer to properties of barnase or its variants (often with respect to particular amino acid positions), or to the effects of amino acid substitutions on barnase, but not to the bacterial host or to the effects of barnase variants on the host cell.

[3] Codon and residue numbering correspond to the sequence of mature wild-type barnase, where the N-terminal residue is Ala. The N-terminal fMet resulting from *synbar* expression is expected to be removed within the cell (*13*), yielding the desired wild-type sequence. Since fMet removal depends on the identity of the adjacent residue, we have left the Ala codon (codon 1) undisturbed.

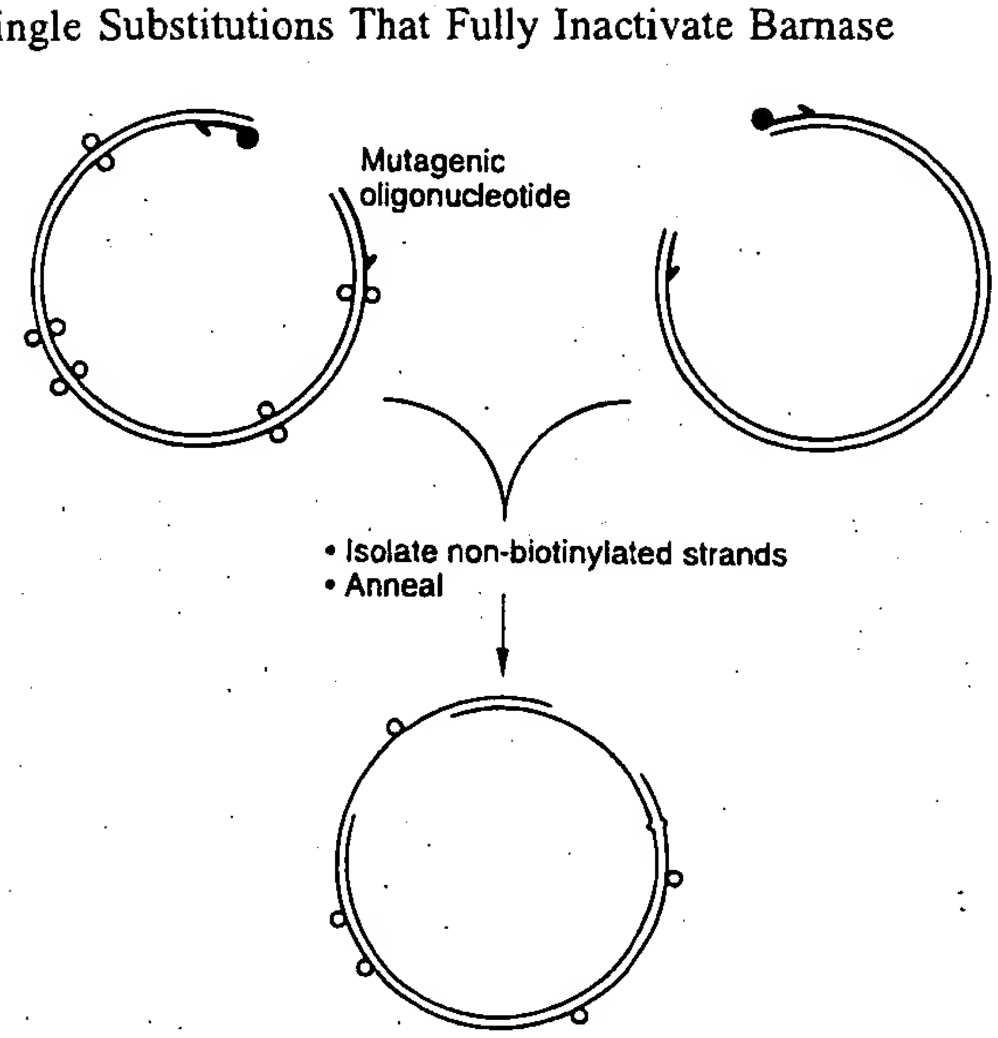


FIGURE 1: Introduction of random base substitutions into *synbar*. The upper illustrations depict the two PCRs used to produce the DNA strands that are combined to form the gapped-duplex product (lower illustration). Plasmid pSNBR serves as the template in both PCRs. Bold half-arrows indicate primers. Filled circles represent biotin groups that are covalently attached to primers. Open circles indicate DNA that has been methylated by Dam methylase. A base substitution in *synbar* results in local mispairing in the final product (as shown).

allow growth of the host cell. Consequently, inactivating mutations are directly selected by plating on Luria-Bertani agar with ampicillin (*12*).

Sequencing of synbar Mutants. Plasmid DNA was prepared from overnight cultures of clones passing selection. Sequences of mutant genes were determined by performing cycle-sequencing reactions with dye-labeled dideoxynucleotides (Perkin-Elmer) and analyzing products with an automated sequencer (Perkin-Elmer, model ABI 373). Clones with multiple substitutions or frame-shift mutations were excluded from our analysis.

Determination of OVA1 Activity. RNase activity has been reported for OVA1, a 15-residue peptide (NVMEERKIKVILPRM) corresponding to a portion of chicken ovalbumin (*15*). To perform a quantitative activity assay, the peptide was synthesized on a commercial synthesizer using FMOC chemistry and purified by HPLC. RNA-hydrolysis activity was determined by measuring the absorbance (301 nm) of a solution consisting of 100 mM Tris buffer (pH 7.6 at 25 °C), 200 mM KCl, 5 mM $MgCl_2$, and 2.8 mg/mL torula yeast RNA (type VI, Sigma), then adding OVA1 (to 30 μM), and monitoring the decrease in absorbance as a function of time. As a reference, a parallel reaction was performed by adding wild-type barnase to the same assay buffer. The relative activity of OVA1 was calculated from the ratio of the absorbance slopes of OVA1 and barnase during the initial steady-state phase of hydrolysis.

RESULTS

Analysis of Completeness. Figure 2 summarizes our findings, indicating all single mutants found to be inactive and all single mutants inferred to be active. Because a limited number of trials were used to identify inactivating substitutions, some such substitutions may, by chance, have escaped detection. Consequently, before the implications of these results are considered, it is important to consider how closely they represent the ideal data set that would result from an infinite number of trials.

One indicator of completeness is the proportion of identified inactivating substitutions for which only one example has been isolated. A very incomplete collection would be dominated by these unduplicated examples, whereas they would become rare as the collection approaches completeness. Our initial plan was therefore to continue the process of collecting and examining mutants until duplicates of all examples had been obtained. However, the highly nonuniform distribution of mutants following selection made it impractical to achieve this.

The problem we encountered is illustrated in Figure 3, which shows all mutants recovered from one of the 8 contiguous *synbar* regions. Codons 83 and 87 normally specify arginine residues. The base mixtures used to prepare the mutagenic oligonucleotide mixture (see Materials and Methods) are expected to give unaltered Arg codons in most plasmid molecules, with a small fraction of plasmids carrying a missense mutation at either of these codons. These mutant codons are expected to specify Cys, Ser, Gly, Leu, Pro, and His with equal frequency. Consequently, if all mutants passing selection were equally inactive, we would recover these inactive mutants at roughly equal frequencies. The fact that frequencies of recovery are highly nonuniform (Figure 3) suggests that significant variation in activity exists even among these inactive mutants.[4] Variation of this sort was sufficiently common in a previous study (*12*) that a third classification was defined for mutants that impaired cell growth without completely preventing it.

Despite varying degrees of inactivity, however, the 64 inactive-mutant isolates represented in Figure 3 clearly demonstrate that positions 83 and 87, where all accessible substitutions lead to inactivation, are considerably more important for barnase function than any of the other positions in the region depicted. While we cannot be certain that an inactivating substitution at position 93, for example, would not be found if another 100 clones carrying mutations in this region were processed, we can be certain that most of these clones would carry substitutions at positions 83 or 87, and we can safely deem it highly unlikely that position 93 is actually as sensitive to substitution as positions 83 and 87 are. We can likewise be confident that positions showing limited sensitivity to substitution (positions 89−91) are less functionally critical than the highly sensitive positions. Moreover, because the identities of the inactivating substitutions found at these three positions are readily explicable in terms of the properties of the introduced side chains (at position 89, normally occupied by a well-buried leucine, Arg is the only accessible substitution that introduces a charged side chain; similarly, Asp is the only charged side chain of those accessible at position 90, normally occupied by a largely buried tyrosine; Pro introduces unusual backbone

[4] The biological interpretation of the observed nonuniformity is that mutants having activities very close to the threshold level may or may not kill the original transformed host cell, depending upon whether the cell is able to make the necessary metabolic adjustments to compensate for the harmful RNase activity. Cells that do make the adjustments form colonies [often smaller than normal colonies (*12*)], but these colonies will be underrepresented to the extent that the survival rates of the initial transformants are reduced.

```
                       D                      R   P       VV
        10         20         30         40         50         60
AQVINTFDGV ADYLQTYHKL PDNYITKSEA QALGWVASKG NLADVAPGKS IGGDIFSNRE

 HFMKSLECF SESMHSSQNF SEKSMSNYDS HSFCCFSPNC KMSEFSSCNP MCCEMLYKCD
 SLFYPIYRL PYHVSPHYSI TYYHFPQQSP SPIRRLPTSR YVPYLTTRST FRRYFIQYSS
 KILHAVHSI THNPKANNQV AHHNLAEKQT KTVSGITAQS HPTHIVASQA LSSHLVKHGQ
 EAVDMSNVA VNDQEIDDEP LNDFVIIEKV EVPVLAVFEV DQVNADLVEF VAANVSEDLK
 LDTIKYVAD DVFRLNFLMH QVICTNTLVE LDHASDDYMA I DVDGQAMY TDDVTYLIPV
 PGNTRCADG GAC P CPTR RAT N RWAG PGRD GGCTD T GAG RDTC N  ANCWTHA
 R  SS G    G  R  RR   GS S   G  R       R  S  G    R  S  GS  S G


                        C   C
                        S   S                 Q
               A        G   G                 Y
               Y        L   L                 N
           S   V        P   P                 D
           CP PH        H   H RD P            R
        70         80         90        100        110
GKLPGKSGRT WREADINYTS GFRNSDRILY SSDWLIYKTT DHYQTFTKIR

CNMSCNRCCS RCDSEMKSSR CL KPE MMS TPECMMSNSS ELSHSLSNMC
RSVTRSCRSP GSQTNFYHPC RI YTY FVH ATYRVFHSPP YPHSPIPSFS
SQPASQGSGA LGKVGLHNAG SV HAH LPN FAHGPLNQAA H NKAVAQLG
VEQLVEIVLI  LVD VDDII VS DFN VQF YFNLQVDEII N DEISIEVL
AMRQAMTAPN  HAG TIFNT AY IYV T C CYVSRTFMNN V FLNYNMTP
DT RDTNDH    G  NTC N DC TCA N    CA  NCT  A CP C TNH
 R   R          SS       S G S     G  S R  G  R   RS
```

FIGURE 2: Summary of single-substitution results for barnase. The wild-type barnase amino acid sequence is shown in bold type, underlined letters indicating residues with side chains that interact directly with substrate (*16*). At each position, substitutions that were found to render the enzyme inactive are shown above the wild-type sequence. All other accessible substitutions (see Materials and Methods) are shown below the wild-type sequence.

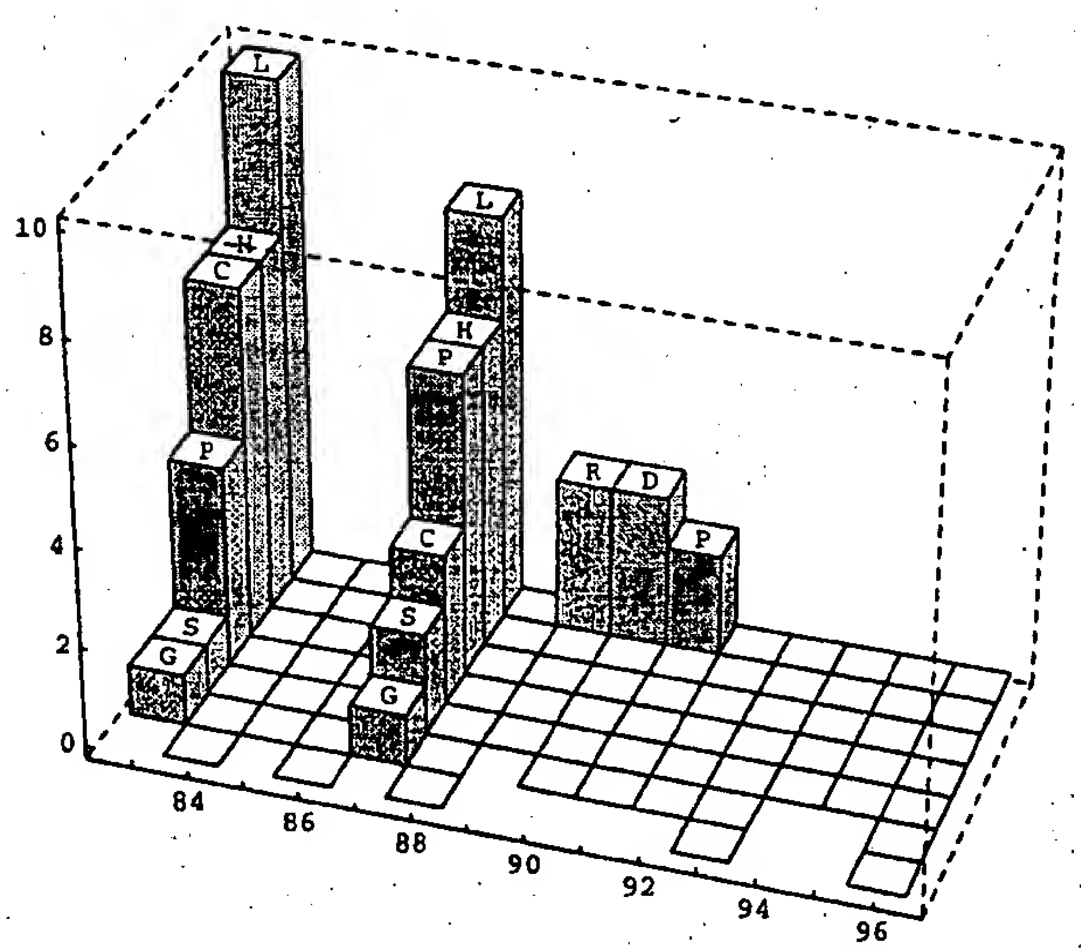


FIGURE 3: Number of independent isolates of inactive barnase variants carrying substitutions from position 83 to position 96. The residue position is indicated on the horizontal axis. Vertical stacks indicate the number (vertical scale) of examples recovered for each accessible substitution in this region. Stacks corresponding to substitutions for which at least one example was recovered are labeled to indicate the introduced amino acid (see Figure 2 for accessible substitutions that were not recovered).

constraints at position 91, normally occupied by a serine situated in the middle strand of a five-strand β-sheet), our approach appears to effectively identify the most severely disruptive substitutions at partially sensitive positions.

Since our primary aim is to identify the residues in wild-type barnase that most critically affect function, the overall substitutional sensitivity of each position is more important than the activity of any particular mutant. Some uncertainty at the level of individual substitutions can be accepted without losing the bigger picture at the level of amino acid positions because several substitutions are possible at each position. In light of this, a more appropriate measure of completeness is the proportion of sensitive positions for which only a single instance of recovering an inactive variant occurred. By this measure, the data set is seen to be complete; multiple examples (averaging 8.6 per position) were isolated for all of the 15 positions found to be vulnerable to inactivating substitution.

Paucity of Inactivating Substitutions. The most striking aspect of the results presented in Figure 2 is the scarcity of inactivating substitutions. Of the 109 positions examined, 94 (86%) appear to tolerate all accessible substitutions, and only two (1.8%) are wholly intolerant of accessible substitutions. As discussed above, it is quite possible that some additional inactivating substitutions would be found if the process of mutant collection and examination were continued indefinitely. If so, this would decrease somewhat the fraction of positions showing complete tolerance. Furthermore, since our method of mutagenesis restricts the substitution set to about 6 substitutions per position, it is highly probable that some inactivating substitutions are inaccessible by this method. However, the overall picture of substitutional tolerance depicted in Figure 2 is not apt to be very different from the true picture (see above). In particular, the number of positions that are wholly intolerant of substitution is more likely to be lower (because of the restricted substitution set) than higher.

DISCUSSION

Extreme tolerance of substitution is not without precedent in studies of this kind. In a study of bacteriophage T4 lysozyme, Rennell et al. (*4*) found that more than half of the positions (55%) tolerate all of the 12 or 13 substitutions tested. More remarkably, only one position of the 163 examined in that study (0.6%) was found to be wholly intolerant of substitution. In another study, Wen et al. found that 75% of the 121 positions in a bacterial membrane protein tolerate nonconservative substitutions (*6*). Although functional tolerance of substitution is clearly an important theme

to emerge from exhaustive-substitution work, counterexamples do exist. The most striking of these, phage P22 Arc repressor, was noted above, as was the importance of the activity threshold in determining the outcome of an exhaustive-substitution study.

A more thorough consideration of the significance of the activity threshold will be instructive at this point. We noted previously that experimental systems used in exhaustive-substitution studies typically allow the experimenter to choose a threshold level from a wide range of feasible values. This raises the question as to whether the distinction between active and inactive is actually arbitrary or whether there is a natural fixed reference point by which these terms might be defined.

Significance of the Activity Threshold

Distinctive Mechanism of Enzymes. Although DNA-binding proteins have been the subject of a number of important exhaustive-substitution studies (*1*, *2*, *7*), we will narrow our focus here to enzymes. This class of molecules catalyzes chemical reactions by doing what all catalysts do, namely, lowering the free energy of the transition-state complex. What distinguishes enzymes from other catalysts is the way in which they accomplish this, and consequently the magnitude of their effect. Unlike simple catalysts, enzymes employ a spatially extensive structure to bind reactants, placing them in a geometrically precise and catalytically optimal orientation (*17*, *18*) relative both to each other (where multiple reactants are involved) and to the catalytic group or groups, which are typically integral to the enzyme.

If separated from their protein scaffold, the catalytic groups alone may catalyze the same reaction in solution by, for example, simple acid or base catalysis. To demonstrate the purpose of the geometric scaffold, however, one need only compare rates of catalysis by small molecules under physiological conditions to rates of enzymatic catalysis for the same reactions. Rates of single-substrate reactions of biological relevance vary by many orders of magnitude when measured in neutral aqueous solutions in the absence of enzymes; first-order rate constants range from 10^{-1} to 10^{-16} s^{-1} for the set of reactions discussed by Radzicka and Wolfenden (*19*). In contrast, effective second-order rate constants (k_{cat}/K_m) for the corresponding enzymatic reactions appear to fall within a relatively narrow range, a typical value being 10^7 s^{-1} M^{-1} (*19*). Using this value and a typical cytoplasmic concentration of 10^{-6} M for an enzyme, we estimate a typical pseudo-first-order rate constant for enzymatic reactions in vivo to be 10 s^{-1}. This can be compared directly to the range of nonenzymatic rate constants given above,[5] indicating that the geometric role of the protein scaffold increases reaction rates by some 2−17 orders of magnitude, depending upon the reaction.

Basal Activity as a Natural Unit of Measure. So universal and so crucial is this geometric aspect of enzymes that it might be viewed as a defining property of this class of molecules. It follows that a natural basal limit to enzyme activity would be a catalytic rate just above that which can be obtained without the spatial positioning employed by enzymes (i.e., the maximal rate of catalysis by a nonenzymatic mechanism defines a limit that can be exceeded only by employing spatial positioning in the manner that is characteristic of enzymes). As discussed above, enzymes typically exceed this basal limit by many orders of magnitude, achieving catalytic perfection in some cases (*19*, *22*, *23*).

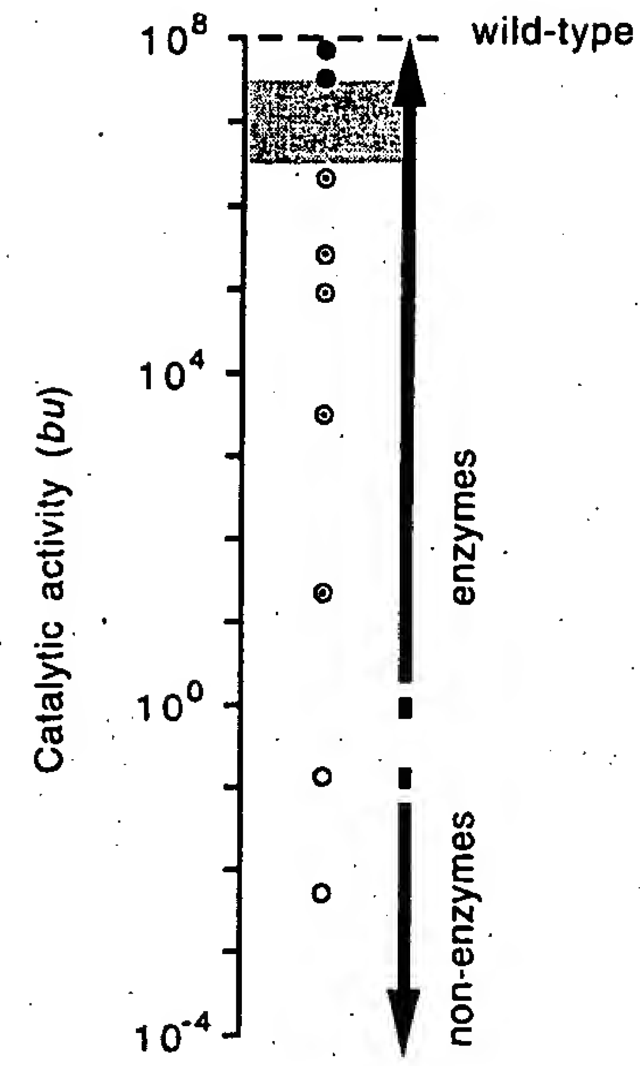


FIGURE 4: Natural scale of activity for a typical enzyme-catalyzed reaction. The scale indicates catalytic activity in terms of basal enzymatic activity units (*bu*), as discussed in the text. In this example, the basal level of activity is 8 orders of magnitude below the activity of the wild-type enzyme. Circles indicate activities of hypothetical mutant enzymes. Open circles correspond to mutants that fail to outperform nonenzyme catalysts and thus function as nonenzymes. All other mutants exceed the basal enzymatic activity level (1 *bu*) and thus function as enzymes. The shaded box indicates a range of activity thresholds from 3 to 30% of wild-type activity. Activity thresholds in the shaded region allow efficient enzymes (filled circles) to be distinguished from less efficient enzymes (circled dots) and nonenzymes, but they do not allow nonenzymes to be distinguished from enzymes.

The scale of catalytic activity shown in Figure 4 uses basal activity as the unit of measure (1 basal enzymatic activity unit, *bu*, corresponds to the basal level of activity described above). The activity of the wild-type enzyme is taken to be 8 orders of magnitude higher than the basal level (i.e., 10^8 *bu*) so a typical enzyme might be represented. Activities of mutants resulting from single amino acid substitutions will span a wide range, from very close to the wild-type level to

[5] Although Radzicka and Wolfenden (*19*) report intrinsic aqueous rate constants (where water is the only catalyst), these values are generally indicative of the level of catalysis that can be expected in micromolar aqueous solutions of nonenzyme solutes. The rate at which small-molecule solutes perform simple acid or base catalysis depends primarily upon their pK_a and concentration (see, for example, the discussion of the Brønsted equation in ref *20*). In neutral aqueous solutions at room temperature (for the purpose of comparing them to enzymes), water is expected to have a more significant catalytic effect than any small solute present at micromolar concentrations, regardless of its pK_a. Like small molecules, macromolecules can act as nonenzyme catalysts, often with the added advantages of substrate binding and local solvent exclusion. For example, Hollfelder et al. have demonstrated that the Kemp elimination is fortuitously catalyzed by serum albumins (*21*). At an albumin concentration of 1 μM, however, their data indicate that catalysis by water is comparable to catalysis by the albumin (pH 8.0 and 25 °C). At micromolar concentrations, then, even these more sophisticated catalysts typically fail to outperform water significantly.

the basal level and lower, depending upon the nature of the change they introduce. The natural significance of 1 *bu* activity, however, provides a meaningful division of this range into two regions; mutants with activities above 1 *bu* continue to exceed the performance of nonenzyme catalysts, whereas mutants with lower activities do not. Thus, even if these less active mutants continue to bind substrate molecules specifically, they fail to bind them in such a way as to enhance catalysis, and they consequently fail to exhibit the characteristic property of enzymes. Conversely, mutants having greater than 1 *bu* activity, however suboptimal they may be, continue to exhibit the characteristic property of enzymes.

Essential versus Refining Structural Features. In this sense, structural features removed upon substitution could be classed as refining features if the activity following substitution exceeds 1 *bu*. Essential features would then be those structural features that, upon removal, reduce activity to less than 1 *bu*.[6] Note that the distinction has significance because of the qualitative difference between enzyme catalysis and nonenzyme catalysis, and not because of the quantitative difference per se. Indeed, for mutants in the vicinity of 1 *bu* activity, the quantitative differences in activities are relatively small. The qualitative distinction, however, remains important: where reversion of single mutants is concerned, the restoration of a refining feature can turn a poor enzyme into a good one, but it cannot turn a nonenzyme into an enzyme.

In exhaustive-substitution studies, the activity threshold effectively divides all single mutants into two classes according to activity (above threshold = active; below threshold = inactive). Although estimates of basal enzymatic activities are not generally made in the course of these studies, in most cases the relative proximity of the threshold to the activity of the wild-type enzyme implies that the threshold is several orders of magnitude above 1 *bu*. For example, in the study of T4 lysozyme by Rennell et al. (*4*), the threshold is placed at 3% of wild-type activity, while in the study of β-lactamase by Huang et al. (*5*), it is placed at about 30% of wild-type activity. This range of threshold levels is indicated in Figure 4. The wild-type activities of these two enzymes may be somewhat higher or lower than the value represented in Figure 4 (10^8 *bu*), but they are not apt to be many orders of magnitude lower. Consequently, we infer that the threshold levels in these studies are very much higher than the basal level of activity. These thresholds, then, would enable the researcher to distinguish refining features of relatively small effect from all other features (both refining and essential), but they would not be useful for distinguishing essential features from refining features.[7] To do this, one would need a system with an activity threshold in the vicinity of 10^0 *bu*.

Essential Features Are Particularly Relevant to Protein Design. In the pursuit of a complete understanding of the relationship between sequence and function, studies of refining features are no less relevant than studies of essential features. Essential features, however, may be of particular importance to efforts in protein design. A reasonable approach to the design of a novel enzyme would be to aim for a rudimentary enzyme as the initial target and then to apply iterative mutation−selection methods (*25−28*) to optimize the crude initial design. Since the sorts of features that are important for the function of natural enzymes will presumably have to be incorporated into any successful designed enzyme, one might hope to facilitate the first stage of the design process by using information from exhaustive-substitution studies of natural proteins to guide the initial design.

One might even be tempted to view an exhaustive-substitution study of a natural enzyme as an exercise aimed at determining the features that would need to be incorporated into a successful re-design of the same enzyme. However, an important limitation of exhaustive-substitution data in this regard is that they cannot be expected to reliably identify unimportant features (i.e., features that would not need to be included in a successful design). The reason for this is that the structural context in which single substitutions are made in an exhaustive-substitution study is that of the wild-type enzyme. Since the thermodynamic stability of natural proteins typically exceeds that which is necessary for function, we would expect there to be many destabilizing substitutions that are functionally tolerated when introduced in the context of the wild-type sequence. When the context is far less optimal, as would be the case for a crude initial design, the same sorts of substitutions might easily turn a weakly functional design into a nonfunctional design.

Can exhaustive-substitution studies be used to identify features that would need to be included in a successful design? In answering this, we will consider refining features and essential features separately. Upon removal of a refining feature, a wild-type enzyme is simply transformed into a suboptimal enzyme. Though suboptimal, this enzyme may still be considerably more active than a rudimentary enzyme of the sort we might hope to create from an initial design. That being the case, it is quite possible that the refining feature in question may only serve as a refinement in the context of other more basic refinements. In a crude enzyme lacking these basic refinements, we cannot assume that this feature would have any significant effect on function. Refining features therefore do not generally provide useful information for the design of rudimentary enzymes.

On the other hand, it seems inescapable that an essential feature of a wild-type enzyme will have a corresponding essential feature in any less optimal variant of that enzyme. That is, if some type of structural modification destroys enzyme function when it was initially optimal, it is difficult to imagine how the same type of modification would not have the same effect when applied to a suboptimal variant. An essential feature, then, points to a structural rule for the class of proteins that perform a particular function by means a particular fold. Exhaustive-substitution data would therfore be of considerable value in obtaining design rules, provided

[6] It is often appropriate to view amino acid substitutions as introducing features as well as (or instead of) removing them. For simplicity of expression, however, we are using the term "feature" in a broad sense to include both the presence and the absence of particular aspects of structure. The absence of a β-carbon at position 52, for example, is a feature of wild-type barnase that is "removed" upon substitution at that position.

[7] A deletion study on the A chain of ricin (*24*) used a sensitive activity test capable of detecting activity down to 0.01% of the wild-type value. However, the extreme specificity of the reaction (cleavage of the N-glycosidic bond of a single adenosine base in the mammalian ribosome) suggests that the rate corresponding to 1 *bu* would be many orders of magnitude lower still.

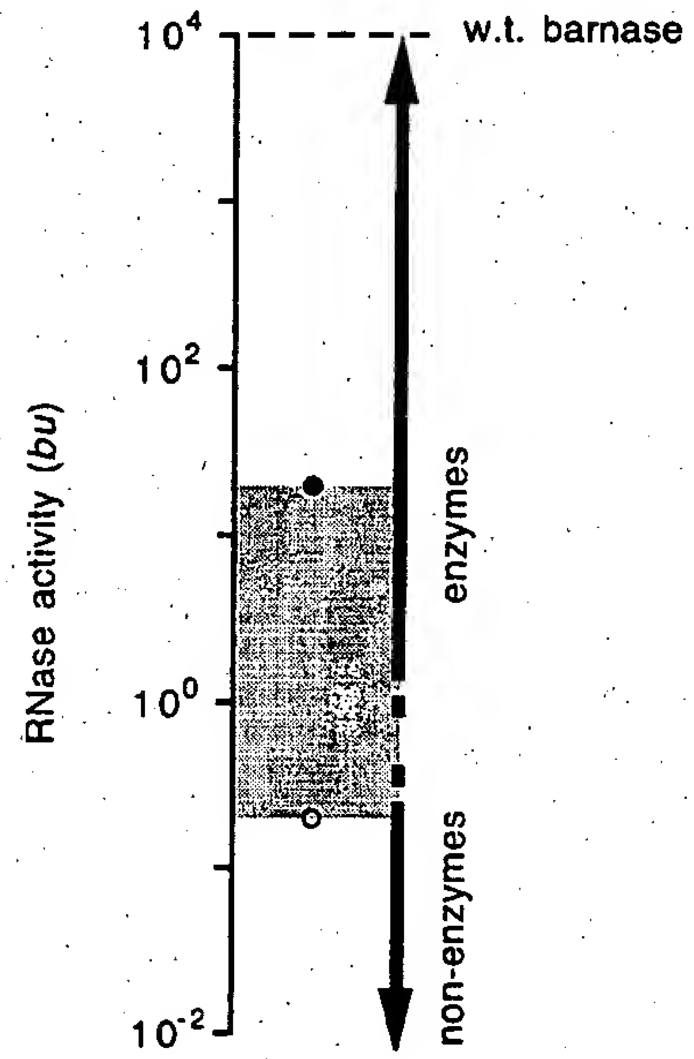


FIGURE 5: Estimation of the activity threshold for the *synbar* selection system. The open circle indicates the RNA-hydrolysis activity of OVA1, a 15-residue peptide corresponding to a portion of chicken ovalbumin (*15*). Because OVA1 has an unusually high activity for a nonenzyme (0.002% of that of wild-type barnase), we take it to represent an optimal or near-optimal nonenzyme catalyst. Taking basal enzymatic activity to be somewhat higher than the activity of OVA1, we define the basal enzymatic activity unit as 1 *bu* ≡ 0.01% of the wild-type activity. On this natural scale, the activity of OVA1 is 2.0×10^{-1} *bu*, and the activity of barnase mutant E73A (filled circle) is 2.0×10^1 *bu* (*29*). As indicated by the shaded region, the activity threshold for the *synbar* selection system lies between these two values.

that a basal activity threshold is used to distinguish essential features from refining ones.

It should be noted, though, that essential features and design rules are different things, in that the former is equivalent to a sequence constraint that applies to a particular protein, whereas the latter is a more general constraint that applies to all possible sequences sharing the fold and function of that protein. While identification of essential features for a particular protein provides valuable information on the design rules for the whole class of proteins, these rules cannot necessarily be deduced from experiments on a single protein. Some amount of interpretation will therefore be necessary for tentative classwide design rules to be inferred from exhaustive substitution data.

The synbar System Approaches a Basal-Activity Selection System. To perform an exhaustive-substitution experiment with a basal-activity threshold, one would need a simple screening or selection procedure capable of detecting activity down to a level that approaches nonenzymatic activity. This presents considerable practical difficulties that may be insurmountable for many systems. The *synbar* selection system described here is one system where the necessary sensitivity appears to be attainable, or nearly so.

Figure 5 depicts the relationship between the *synbar* threshold and known enzymatic and nonenzymatic rates of RNA hydrolysis. The barnase mutant E73A is nearly 3 orders of magnitude less active than the wild-type enzyme because it lacks the side chain that normally acts as the catalytic general base in the first step of the hydrolysis reaction (*30*). Since this mutant tests active in the *synbar* system (*12*), the activity threshold for this system must lie below the activity of the mutant, as indicated in Figure 5. As a lower bound to the selection threshold, we will consider a particular class of peptide catalysts.

Yanagawa and co-workers have demonstrated that some peptide fragments of barnase catalyze RNA hydrolysis (*15*). Although they drew the conclusion that this activity is relevant to the function of the whole enzyme, their demonstration that completely unrelated peptides show the same activity undermines that conclusion. Their further demonstration that for peptides to exhibit this activity they need only carry a net charge of +2 or more argues convincingly that the activity they have observed is not enzymatic in nature. However, with activities approximately 5 orders of magnitude below that of wild-type barnase, these are remarkably active catalysts for nonenzymes. Their ability to bind RNA (*15*) accounts, at least in part, for their catalytic performance.

We have synthesized one of these peptide catalysts (see Materials and Methods) and determined its activity to be 0.002% of that of wild-type barnase (mole-to-mole basis). Taking this to be an approximate upper-limit rate of nonenzymatic RNA hydrolysis under physiological conditions, we estimate basal enzymatic activity to be approximately 4 orders of magnitude below the activity of wild-type barnase. This defines the basal enzymatic activity unit used in the scale of Figure 5. Under the reasonable assumption that barnase mutants must outperform nonenzyme catalysts in order to exhibit lethality (evidence that lethality requires essentially barnase-like structure, and hence proper enzyme function, is discussed below), we conclude that the selection threshold for the *synbar* system must lie above 0.2 *bu*. Situated between 2.0×10^{-1} and 2.0×10^1 *bu*, then, the threshold can reasonably be said to be in the vicinity of the basal level, 1×10^0 *bu*. Consequently, substitutions that lead to activities below the threshold (i.e., to nonlethal barnase variants) must reduce activity to, or nearly to, nonenzymatic levels.

Inactivating Mutations in Perspective

Inspection of the collection of mutants having this dramatic effect suggests that they fall into three classes (Table 1). The first of these, class I, includes all substitutions that replace a side chain known to be directly involved in substrate binding or catalysis. The 17 substitutions falling into this class all involve replacement of Arg83, Arg87, or His102. As with all substitutions, there is a possibility of the local change at the site of substitution leading to a more extensive structural disturbance. However, because of the crucial and direct role of these three residues in function, such propagated disturbances would not need to be present to account for the effects of substitution. Therefore, class I substitutions will be excluded from subsequent classes, even if they might otherwise meet the criteria for inclusion.

Class II includes all substitutions (not in class I) that replace a side chain that is substantially buried (i.e., <10% solvent-exposed) in the wild-type structure. Although the number of mutants falling into this class is similar to the number in the previous class, the number of positions involved is significantly greater. Consequently, positions that contribute to this class tend to be considerably less vulnerable

Table 1: Classification of Inactivating Substitutions

WT residue	solvent exposure of the WT residue (%)[a]	class I	class II	class III
Tyr24	9 (1)	–	D	–
Leu42	0 (0)	–	R	–
Ala46	2 (0)	–	P	P
Gly52	4	–	V	V
Gly53	27	–	–	V
Trp71	2 (2)	–	C, S	–
Arg72	23 (28)	–	–	P
Ala74	0 (0)	–	P	P
Asp75	0 (0)	–	A, H, Y, V	–
Arg83		C, G, H, L, P, S	–	–
Arg87		C, G, H, L, P, S	–	–
Leu89	0 (0)	–	R	–
Tyr90	7 (7)	–	D	–
Ser91	1 (0)	–	P	P
His102		D, N, Q, R, Y	–	–

[a] Solvent-accessible surface areas, calculated by the method of Lee and Richards (*31*), are given as percentages of the areas of each amino acid X in an extended Gly-X-Gly tripeptide (*32*). The first value for each position applies to the entire residue; values in parentheses apply to side chains alone (omitted for glycines). Exposure values are not given for residues that contact substrate because inactivating substitutions at these positions are exclusively assigned to class I.

to inactivating substitution (i.e., only a small fraction of the accessible substitutions destroy function) than the positions involved in class I substitutions. Position 75, where 4 of 7 possible substitutions are inactivating, provides a possible exception to this. In the wild-type enzyme, the aspartate side chain at this position forms a salt bridge with the arginine side chain at position 83, one of the 3 positions that account for all class I substitutions. The extreme sensitivity to modification at position 83 suggests that the sensitivity exhibited at position 75 might be due to the close interaction between these two residues.

Many of the substitutions in class II involve replacement of a hydrophobic side chain with a polar or charged one. The exceptions to this, however, are sufficiently numerous that they suggest another cause of inactivation. Thus, class III includes all substitutions (not in class I) that either introduce a proline residue or replace a glycine residue. Because the proline side chain places unusual restrictions on backbone conformation and the glycine side chain does just the opposite, both types of substitution have a strong tendency to introduce local backbone distortion. As shown in Table 1, 6 substitutions can be placed into class III, several of them also falling into class II. Since there are 66 positions where Pro is not an accessible substitution (see Figure 2), the true size of class III is probably somewhat larger.

It is noteworthy that these three classes give a complete description of the kinds of single substitutions that can destroy this enzyme. Considering the level of functional impairment required by *synbar* selection, we would expect the disruptive effects of these sorts of substitutions to be evident from previous work on other proteins. This is indeed the case. Earlier exhaustive-substitution studies, for example, have demonstrated the functional importance of particular active-site residues (*4*–*6*), the higher functional sensitivity at buried positions (*1*, *4*, *5*, *7*), the unusually disruptive nature of proline (*1*), and the unusual sensitivity to substitution of particular glycine residues (*3*, *4*). What the *synbar* system reveals is the extent to which the corresponding structural features (active-site groups, buried side chains, and local backbone conformation) are essential to enzyme function at the most basic level.

Four primary conclusions are evident in this regard. First, it is a rare single substitution that is capable of destroying barnase function (i.e., reducing it to a level that can be approached by nonenzymes). Only about 5% of the accessible substitutions in this study were found to have an effect so severe. Second, in all cases where a substitution does have this effect, the cause (in broad terms) appears to be either (a) direct modification of the active site, (b) nonconservative replacement of a buried side chain, or (c) introduction of non-native local backbone constraints. Third, even among substitutions falling into these categories, elimination of enzyme function is atypical. For example, of the eight positions where side chains interact directly with substrate (indicated in Figure 2), only three are vulnerable to inactivating substitution, and only two of those appear to be wholly intolerant of substitution. Finally, all of the wild-type residues that exhibit extreme sensitivity to substitution (i.e., where a substantial majority of the accessible substitutions eliminate enzyme function) interact directly with substrate.

Having surveyed the set of inactivating substitutions, we can examine an important point that was presumed to be true in the previous section, namely, that barnase mutants must function as enzymes to exhibit lethality in the *synbar* selection system. The mere fact that some single substitutions can render barnase nonlethal in this system suggests that proper enzymatic activity [as opposed to the hydrolytic activity exhibited by peptides such as OVA1 (Figure 5)] is required for lethality. The fact that all instances of inactivation can be explained with reference to particular aspects of the structure and mechanism of wild-type barnase strengthens this conclusion because it indicates that these aspects are necessary for lethality. The conclusion becomes even more compelling when we consider what the inactivating substitutions tell us about the role of larger structural elements in forming a lethal protein.

Of the 15 positions that are sensitive to substitution, 12 are not involved in any direct interaction with substrate. Inactivation by substitution at these 12 positions must therefore result from propagated structural disturbances. As shown in Figure 6, these points of structural sensitivity are distributed among four elements of secondary structure (the third of three helices and the first three of five β-strands). Since the enzyme can be rendered nonlethal by propagated structural changes arising from substitutions in these structural elements, we can safely conclude that these elements must be at least partly formed for lethality to be possible. By preparing and testing truncated mutants lacking either the major α-helix (helix 1) or the final β-strand, we have further determined that these two elements must be present for lethality to be possible (D. D. Axe, unpublished result). This implies that all five strands of the sheet must be in place (the cooperative nature of β-sheets makes it implausible that a single internal strand could be unformed). Thus, of the eight elements of secondary structure in barnase, seven must be at least partly formed for a mutant to be lethal. Taken together with the three sensitive active-site positions, this means that the molecule must be largely intact for the host cell to be killed by it, and this confirms our earlier presumption that mutants scored as active are true enzymes.

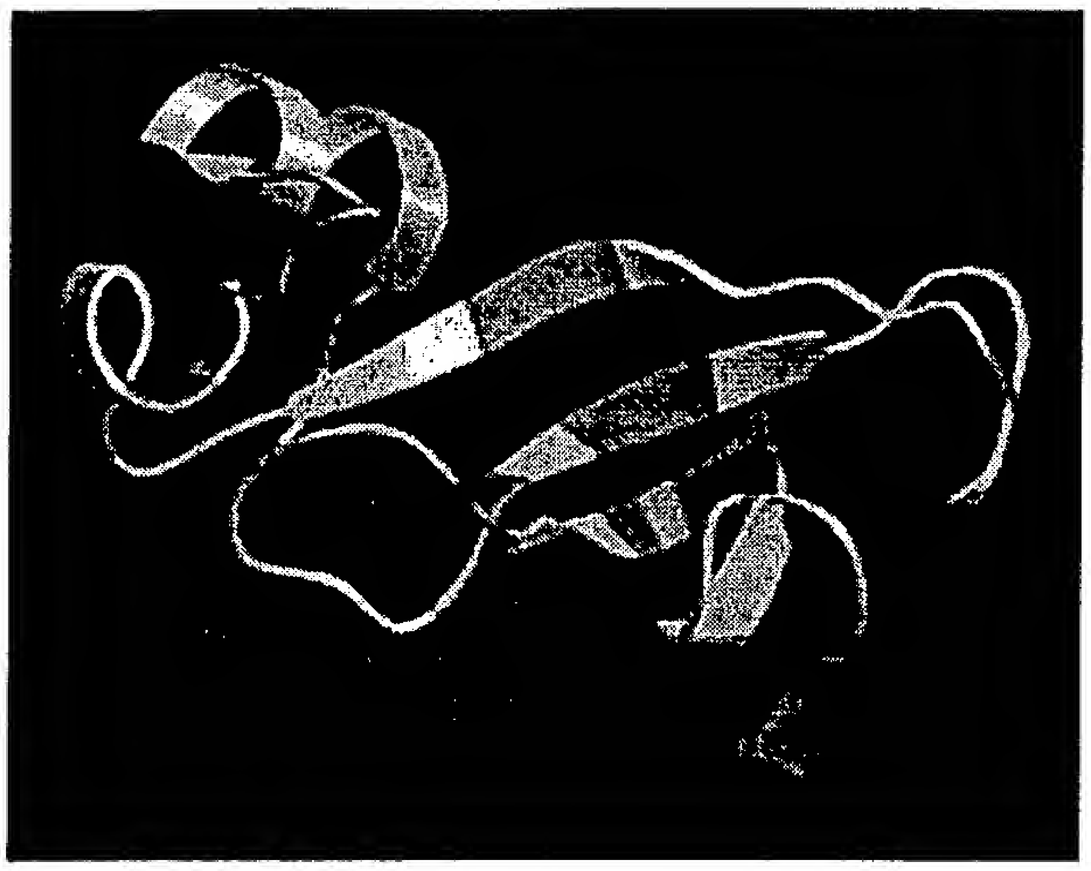


FIGURE 6: Location in the barnase structure of the 15 positions found to be vulnerable to inactivating substitution. Orange indicates the 3 positions where the side chain interacts directly with substrate in the wild-type enzyme (*16*). The other 12 sensitive positions are colored magenta. In addition, activity has been shown to be eliminated by both N-terminal and C-terminal truncations. The missing portions in these inactive constructs are shown in blue and green, respectively.

Issues Calling for Further Study

A number of interesting questions are raised by the results of this work. Class I is probably of more interest for what it does not contain than for what it does. That either of the active-site residues E73 or H102 [normally filling the roles of catalytic general acid and base in the two-step reaction (*30*)] can be replaced without destroying the enzyme raises the interesting question of how the enzyme compensates for their absence. Of particular interest among class II substitutions are those where the introduced side chain is not highly hydrophilic and the substitution does not fall into class III. W71C is the best example of such a substitution.

This tryptophan side chain normally packs against a cluster of other hydrophobic side chains to form a small hydrophobic core (*33*). Evidently, the structural shifts that result when a much smaller side chain is substituted can dramatically impair function. Although arginine is far more hydrophilic than cysteine, our inability to recover a W71R mutant suggests that the hydrophobic portion of the arginine side chain is a better tryptophan substitute in this case than the cysteine side chain is. Since W71C is inactive, one would have thought that W71G would also be inactive. As discussed above, we cannot conclusively declare a mutant to be active on the basis of our inability to recover it. It is possible, then, that W71G is inactive despite the fact that it was not recovered (Figure 2). The best way to conclusively verify that a particular mutant is active is to prepare the appropriate mutant plasmid and apply the *synbar* test as a screen (as in ref *12*).

Among the interesting class III substitutions are the ones that replace either of the two glycines at positions 52 and 53. At both positions, valine is seen to be inactivating, but a number of nonconservative substitutions apparently do not completely eliminate activity. The fact that numerous independent examples of the valine substitutions were recovered (seven examples of G52V and four examples of G53V) suggests that most of the mutants that were not recovered really are active. The presence of a β-bulge (a β-sheet distortion caused by a surplus residue in one strand) involving residues 53 and 54 (*34*) raises the interesting possibility that this small structural element may be important for barnase function. The results of a detailed investigation of the effects of substitutions in this region will be reported elsewhere.

Design Implications

In light of the above discussion on the significance of essential features, we should now consider what the results of this study imply for the design of a barnase-like enzyme. For the reasons indicated previously, we must here focus on those features that appear to be essential, recognizing that the list of these will probably be incomplete. The first implication to consider is that any protein design aspiring to emulate the fold and mechanism of barnase will need two arginines to fill the roles of R83 and R87, and probably a histidine to fill the role of H102 as well (H102 does not appear to be truly essential, but it is sufficiently sensitive to substitution to suggest that it may be essential in anything but a highly optimal context). As discussed above, though, essential features of wild-type barnase cannot generally be construed as design rules for barnase-like enzymes. Since the experiment described here looks at single mutants only, it does not rule out the possibility that one or more of these active-site residues might be replaceable in the context of appropriate compensating substitutions. If we view this work more broadly, though, as giving us a picture of what kinds of single-residue features are indispensable in the context of a natural enzyme, we conclude that a barnase-like enzyme with barnase-like activity will have a few indispensable active-site residues, their exact identities possibly varying for various designs. A less optimal design, striving only for basal enzymatic activity, would at the very least need to have these few residues in their proper spatial orientations.

Conceivably, this is essentially all that is needed for an enzyme to have basal activity, the trick being to design a scaffold that holds the few key residues in their proper orientations. That is, it is reasonable to view direct interaction with substrate as a prerequisite for a residue to be considered to have a direct role in function. The role of the remaining residues, the scaffold residues, is then to impart the necessary orientations, structural dynamics, and chemical properties to the residues on the "front line". Again, though, experiments with single substitutions cannot be expected to give a full picture of the complexity of this front line. It may well be that some of the barnase residues known to interact directly with substrate but found here not to be vulnerable to inactivating substitutions (e.g., E73) would be essential in a less optimal context. What we conclude from this study, then, is that none of the scaffold residues are irreplaceable in the context of an otherwise wild-type sequence (even D75, the scaffold residue found to be most sensitive to substitution, can be replaced without eliminating activity). Studies involving combined substitutions (as in ref *12*) will provide a clearer picture of the sequence requirements for basal barnase function. This work constitutes an essential first step, as it provides information that is needed for the design of multiple-substitution experiments.

ACKNOWLEDGMENT

We thank Dr. Wai Chen for his assistence in the preparation of Figure 6 and Dr. Brian DeDecker for his assistence in the calculation of solvent-accessible surface areas.

REFERENCES

1. Markiewicz, P., Kleina, L. G., Cruz, C., Ehret, S., and Miller, J. H. (1994) *J. Mol. Biol. 240*, 421–433.
2. Bowie, J. U., and Sauer, R. T. (1989) *Proc. Natl. Acad. Sci. U.S.A. 86*, 2152–2156.
3. Loeb, D. D., Swanstrom, R., Everitt, L., Manchester, M., Stamper, S. E., and Hutchison, C. A. (1989) *Nature 340*, 397–400.
4. Rennell, D., Bouvier, S. E., Hardy, L. W., and Poteete, A. R. (1991) *J. Mol. Biol. 222*, 67–87.
5. Huang, W., Petrosino, J., Hirsch, M., Shenkin, P. S., and Palzkill, T. (1996) *J. Mol. Biol. 258*, 688–703.
6. Wen, J., Chen, X., and Bowie, J. U. (1996) *Nat. Struct. Biol. 3*, 141–148.
7. Terwilliger, T. C., Zabin, H. B., Horvath, M. B., Sandberg, W. S., and Schlunk, P. M. (1994) *J. Mol. Biol. 236*, 556–571.
8. Huang, X., and Boxer, S. G. (1994) *Nat. Struct. Biol. 1*, 226–229.
9. Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A., and Sauer, R. T. (1990) *Science 247*, 1306–1310.
10. Suckow, J., Markiewicz, P., Kleina, L. G., Miller, J., Kisters-Woike, B., and Muller-Hill, B. (1996) *J. Mol. Biol. 261*, 509–523.
11. Hartley, R. W. (1988) *J. Mol. Biol. 202*, 913–915.
12. Axe, D. D., Foster, N. W., and Fersht, A. R. (1996) *Proc. Natl. Acad. Sci. U.S.A. 93*, 5590–5594.
13. Hirel, P. H., Schmitter, J. M., Dessen, P., Fayat, G., and Blanquet, S. (1989) *Proc. Natl. Acad. Sci. U.S.A. 86*, 8247–8251.
14. Modrich, P. (1991) *Annu. Rev. Genet. 25*, 229–253.
15. Yanagawa, H., Yoshida, K., Torigoe, C., Park, J. S., Sato, K., Shirai, T., and Go, M. (1993) *J. Biol. Chem. 268*, 5861–5865.
16. Buckle, A. M., and Fersht, A. R. (1994) *Biochemistry 33*, 1644–1653.
17. Knowles, J. R. (1991) *Nature 350*, 121–124.
18. Mesecar, A. D., Stoddard, B. L., and Koshland, D. E. (1997) *Science 277*, 202–206.
19. Radzicka, A., and Wolfenden, R. (1995) *Science 267*, 90–93.
20. Fersht, A. R. (1985) *Enzyme Structure and Mechanism*, 2nd ed., W. H. Freeman, New York.
21. Hollfelder, F., Kirby, A. J., and Tawfik, D. S. (1996) *Nature 383*, 60–63.
22. Blacklow, S. C., Raines, R. T., Lim, W. A., Zamore, P. D., and Knowles, J. R. (1988) *Biochemistry 27*, 1158–1167.
23. Ellerby, L. M., Cabelli, D. E., Graden, J. A., and Valentine, J. S. (1996) *J. Am. Chem. Soc. 118*, 6556–6561.
24. Morris, K. N., and Wool, I. G. (1992) *Proc. Natl. Acad. Sci. U.S.A. 89*, 4869–4873.
25. Stemmer, W. P. C. (1994) *Nature 370*, 389–391.
26. Wright, M. C., and Joyce, G. F. (1997) *Science 276*, 614–617.
27. Tarasow, T. M., Tarasow, S. L., and Eaton, B. E. (1997) *Nature 389*, 54–57.
28. Zhang, B., and Chech, T. R. (1997) *Nature 390*, 96–100.
29. Mossakowska, D. E., Nyberg, K., and Fersht, A. R. (1989) *Biochemistry 28*, 3843–3850.
30. Day, A. G., Parsonage, D., Ebel, S., Brown, T., and Fersht, A. R. (1992) *Biochemistry 31*, 6390–6395.
31. Lee, B., and Richards, F. M. (1971) *J. Mol. Biol. 55*, 379–400.
32. Miller, S., Janin, J., Lesk, A. M., and Chothia, C. (1987) *J. Mol. Biol. 196*, 641–656.
33. Serrano, L., Kellis, J. T., Cann, P., Matouschek, A., and Fersht, A. R. (1992) *J. Mol. Biol. 224*, 783–804.
34. Chan, A. W. E., Hutchinson, E. G., Harris, D., and Thornton, J. M. (1993) *Protein Sci. 2*, 1574–1590.

BI9804028

J. Mol. Biol. (1991) 222, 67–87

Systematic Mutation of Bacteriophage T4 Lysozyme

Dale Rennell, Suzanne E. Bouvier, Larry W. Hardy†
and Anthony R. Poteete‡

Department of Molecular Genetics and Microbiology
and Program in Molecular Medicine
University of Massachusetts, Worcester, MA, U.S.A.

(Received 4 April 1991; accepted 9 July 1991)

Amber mutations were introduced into every codon (except the initiating AUG) of the bacteriophage T4 lysozyme gene. The amber alleles were introduced into a bacteriophage P22 hybrid, called P22 e416, in which the normal P22 lysozyme gene is replaced by its T4 homologue, and which consequently depends upon T4 lysozyme for its ability to form a plaque. The resulting amber mutants were tested for plaque formation on amber suppressor strains of *Salmonella typhimurium*. Experiments with other hybrid phages engineered to produce different amounts of wild-type T4 lysozyme have shown that, to score as deleterious, a mutation must reduce lysozyme activity to less than 3% of that produced by wild-type P22 e416. Plating the collection of amber mutants covering 163 of the 164 codons of T4 lysozyme, on 13 suppressor strains that each insert a different amino acid residue in response to the amber codon, tests the effects of multiple single amino acid substitutions at every position in the protein (except the first). Of the resulting 2015 single amino acid substitutions in T4 lysozyme, 328 were found to be sufficiently deleterious to inhibit plaque formation. More than half (55%) of the positions in the protein tolerated all substitutions examined. Among (N-terminal) amber fragments, only those of 161 or more residues are active.

The effects of many of the deleterious substitutions are interpretable in light of the known structure of T4 lysozyme. Residues in the molecule that are refractory to replacements generally have solvent-inaccessible side-chains; the catalytic Glu11 and Asp20 residues are notable exceptions. Especially sensitive sites include residues involved in buried salt bridges near the catalytic site (Asp10, Arg145 and Arg148) and a few others that may have critical structural roles (Gly30, Trp138 and Tyr161).

Keywords: amber mutations; single amino acid substitutions; critical residues

1. Introduction

Bacteriophage T4 lysozyme mutants have been useful in studies of basic questions in molecular biology. Combined genetic and protein sequencing studies of T4 lysozyme mutants were significant in securing our present understanding of the genetic code (see Streisinger *et al.*, 1966). More recently, combined genetic and structural studies of this protein have yielded insights into the structural determinants of protein stability (for a review, see Matthews, 1987). T4 lysozyme is an especially suitable protein for structural studies. The structure of the wild-type enzyme has been determined crystallographically and refined to a resolution of 1·7 Å (1 Å = 0·1 nm; Remington *et al.*, 1978; Weaver & Matthews, 1987). Moreover, many mutant variants of T4 lysozyme have been found to form isomorphous crystals of high quality; it has been possible to examine closely the structural effects of numerous amino acid substitutions in this protein (see Alber & Matthews, 1987). T4 lysozyme is a likely object for studies aimed at uncovering the sequence determinants of protein structure, a subject sometimes called the second half of the genetic code.

Previous studies of mutant T4 lysozymes have generally focused on small numbers of mutations. Cumulatively, many mutants have been described, but the relative contributions to biological function of many residues in the molecule are unknown. A systematic survey of the effects of single amino acid

† Present address: Department of Pharmacology and Program in Molecular Medicine, University of Massachusetts, Worcester, MA, U.S.A.

‡ Author to whom correspondence should be addressed at: Department of Molecular Genetics and Microbiology, University of Massachusetts Medical Center, 55 Lake Ave N., Worcester, MA 01655, U.S.A.

0022-2836/91/210067-22 $03.00/0

substitutions in T4 lysozyme would generate a functional map, which could be informative when correlated with the structure. Ideally, such a survey would include all 19 single amino acid substitutions at every position, but even for a protein as small as T4 lysozyme (164 amino acid residues: Tsugita & Inouye, 1968), over 3000 variants would have to be studied. To reduce the scope of the technical undertaking, we employed the approach used by Miller and co-workers in studies of the *Escherichia coli lac* repressor (Miller *et al.*, 1979; Kleina & Miller, 1990). In this approach, amber mutations are introduced into the gene encoding the protein in question, and the resulting mutant allele is tested in suppressor strains that insert different amino acids in response to the amber codon. In this way, it is possible, with each single mutant allele of the gene, to test a number of single amino acid substitutions equal to the number of available amber suppressor specificities.

We have introduced amber mutations into all but the first codon of the T4 lysozyme gene. The resulting set of 163 mutant phages has been plated on a set of 13 amber suppressor strains, each of which inserts a different amino acid. The suppressors employed in these studies include four naturally occurring (Winston *et al.*, 1979) and nine synthetic amber suppressors (Normanly *et al.*, 1986, 1990; Kleina *et al.*, 1990; McClain & Foss, 1988). These platings permit a screening of the functional effects of multiple single amino acid substitutions at every position in a protein of known structure. In a related study, Loeb *et al.* (1989) tested the effects of one to ten different amino acid substitutions at every position in the HIV-1 protease. The effects of a great number of amino acid substitutions in the *E. coli lac* repressor have been reported (Miller *et al.*, 1979; Kleina & Miller, 1990); however, the structure of this protein has not been determined crystallographically. An N-terminal fragment of λ *cI* repressor (Reidhaar-Olson & Sauer, 1988; Bowie *et al.*, 1990) has been studied by combinatorial cassette mutagenesis. This procedure produces a related data base, namely, which combinations of amino acid substitutions are tolerated by a protein. We compare the trends in the data presented here for T4 lysozyme with findings derived from these similarly comprehensive collections of mutants.

2. Materials and Methods

(a) *Media, enzymes and buffers*

LB broth contained 10 g tryptone, 5 g yeast extract, 5 g NaCl and 1 ml 1 M-NaOH/l; LB agar contained, in addition, 11 g agar/l. It was supplemented with tetracycline at 5 to 15 μg/ml, or ampicillin at 50 to 100 μg/ml where appropriate. Lambda agar contained 10 g tryptone, 2·5 g NaCl and 11 g agar/l; top agar was identical, except for containing 9 g agar/l. M9 minimal medium was 42 mM-Na_2HPO_4, 20 mM-KH_2PO_4, 8·5 mM-NaCl, 18·7 mM-NH_4Cl, 1 mM-$MgSO_4$, 50 μg thiamine/ml and 4 mg glucose/ml; minimal agar contained, in addition, 15 g agar/l. Where indicated, leucine was added at 40 μg/ml.

Restriction buffers, kinase buffer, ligase buffer and polymerase buffers were as recommended by the suppliers. SB was 0·4 M-Na_2HPO_4, 2·2 M-KH_2PO_4; BS (buffered saline) was made by mixing 9 vol. 0·85% (w/v) NaCl with 1 vol. SB. 5×AB was 250 mM-Tris·HCl (pH 8·3), 300 mM-NaCl, 50 mM-dithiothreitol. 5×RT was 250 mM-Tris·HCl (pH 8·3), 300 mM-NaCl, 50 mM-dithiothreitol, 150 mM-magnesium acetate. Each ddNTP mixture was made with 3 μl of 2·5 mM-ddNTP, 10 μl of 5×AB and 37 μl of water. The dNTP mixture contained each dNTP at 2 mM in AB buffer. Reverse transcriptase mixture contained 4·5 μl reverse transcriptase (Promega; 7 units/μl) in 40 μl of RT buffer. Sequencing gels and related buffers were as described by Sambrook *et al.* (1989).

T7 DNA polymerase and restriction enzyme *Hpa*I were purchased from Boehringer-Mannheim. T4 DNA ligase, Klenow fragment of *E. coli* DNA polymerase I, T4 polynucleotide kinase, and other restriction enzymes were purchased from New England Biolabs. Low-temperature gelling agarose (SeaPlaque) was purchased from FMC Bioproducts; [γ-^{32}P]ATP (6000 Ci/mmol) was purchased from New England Nuclear.

(b) *Plasmids*

The amber suppressor-bearing plasmids pGFIB:sup GLY, pGFIB:sup ALA, pGFIB:sup HIS A, pGFIB:sup GLU, pGFIB:sup LYS and pGFIB:sup PRO H (Normanly *et al.*, 1990) were obtained from J. Miller; pFTOR1 Δ26 (ARG) (McClain & Foss, 1988) was obtained from W. McClain; pDR463 and pDR464 were derived from pGFIB:Phe and pGFIB:Cys (Normanly *et al.*, 1986), and have been described (Rennell & Poteete, 1989). Plasmid pZ152 (Zagursky & Berman, 1984) was used as a cloning vector.

Plasmid pLH416 (Hardy & Poteete, 1991; see Fig. 1) was used as the target for introduction of amber mutations into the T4 lysozyme gene (*e*) by oligonucleotide-directed mutagenesis. This plasmid contains the *bla* gene, origin of replication and filamentous phage IG sequence from pZ152, and a modified fragment of P22 DNA extending from the *Rsa*I site in gene *13* to the *Hpa*I site in gene *15*. The P22 segment was altered by the replacement of sequences between the *Hinf*I site in the 5′ end of gene *19* and codon 54 of gene *15* with: (1) a synthetic DNA segment containing a stop codon for gene *19* and an *Nde*I site overlapping the initiating ATG codon of T4 gene *e*; and (2) a segment of T4 DNA containing the rest of gene *e* and 42 bases downstream from it. The *Nde*I site normally present in the pZ152 sequences was removed by cutting with *Nde*I, filling in and religating, so that the *Nde*I site overlapping the translational initiation codon of gene *e* would be unique.

Plasmid constructions were done by using standard methods (Sambrook *et al.*, 1989). A segment of DNA from pBR322 containing the tetracycline resistance determinant was inserted into pGFIB:sup GLY and pGFIB:sup ALA, generating pDR621 and pDR622, respectively. This was done by digesting the plasmids with *Cla*I, filling in the ends, and ligating with the smaller fragment of pBR322 generated by digesting with *Eco*RI and *Pvu*II and filling in the ends.

A series of plasmids containing small portions of the T4 lysozyme gene was constructed for marker rescue experiments (see Fig. 1). Plasmid pDR609, which contains codons 47 to 164, was made by digesting pLH449 with *Sac*I and *Sal*I, filling in the ends and religating; pLH449

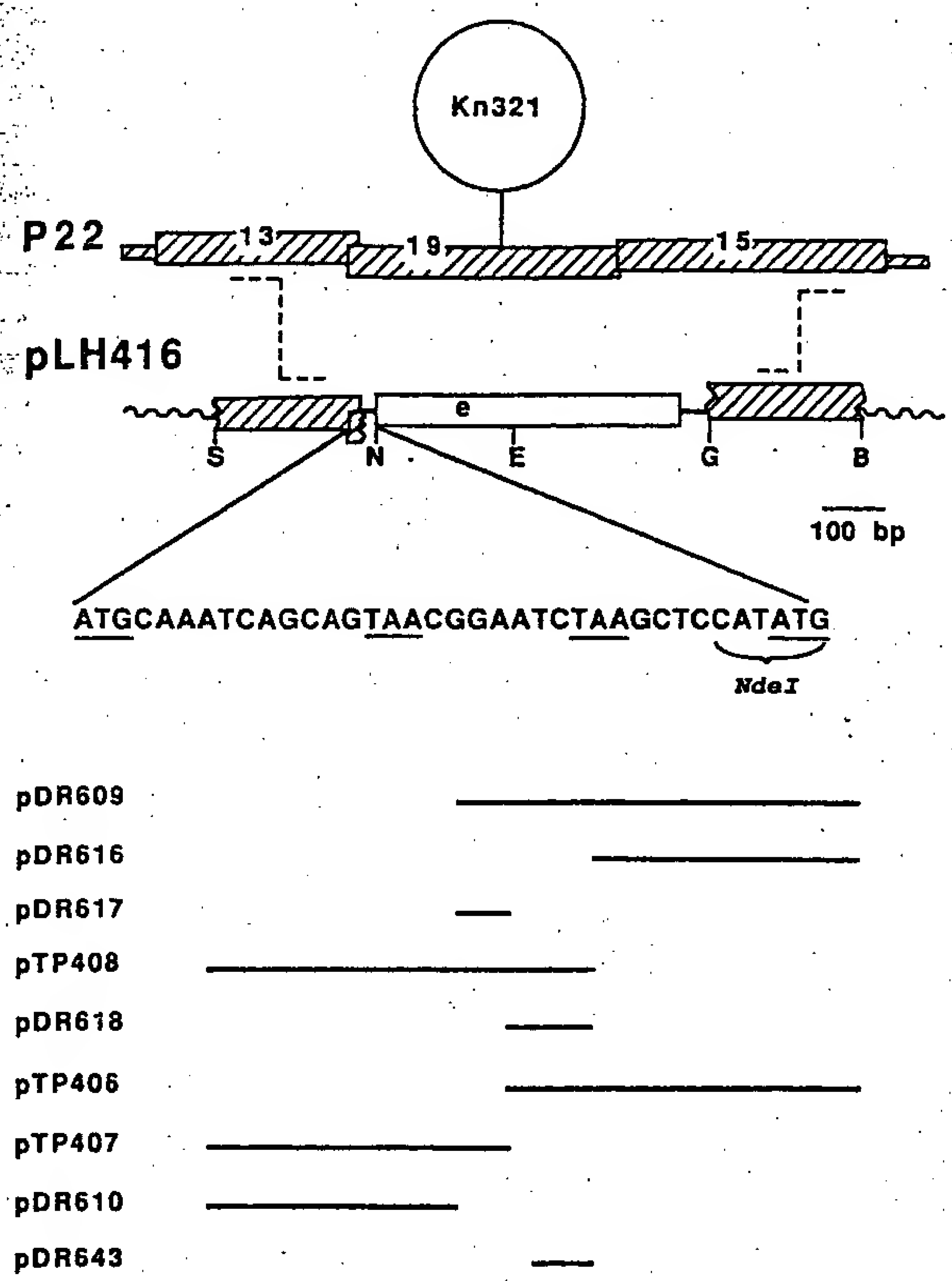


Figure 1. Genetic structure of P22 *e416*. The top line shows the structure of the P22 chromosome in the vicinity of its lysozyme gene (*19*) with the location of the *Kn321* insertion indicated. The corresponding part of plasmid pLH416 is drawn below, with part of its DNA sequence near the start of gene *e*. The 4 underlined codons are (left to right): the translational initiation codon of gene *19*, the stop codon of gene *13*, a synthetic stop codon for gene *19*, and the translational initiation codon of gene *e*. The broken lines indicate crossover points that lead to the generation of P22 *e416* by homologous recombination between the defective prophage P22 *Kn321 sieA44 m44* and pLH416. The hybrid phage lacks P22 gene *19*, which is replaced by T4 gene *e*. It also lacks gene *15* function; however, this gene contributes relatively little to the plaque-forming ability of P22 under the conditions employed in this study (Casjens *et al.*, 1989). The sequence of the *e*416 substitution near the 5′ end of gene *e* is indicated. Segments of gene *e* borne by plasmids used for marker rescue experiments are indicated below. bp, base-pairs.

was generated by primer-directed mutagenesis of pLH416, changing bases 135 to 138 of gene *e* from ATTA to GCTC, creating a *Sac*I site without changing the amino acid sequence of the lysozyme. Plasmid pDR610 was generated by digesting pLH449 with *Sac*I and *Bam*HI, filling in and religating; it contains codons 1 to 44. Plasmid pDR616 was generated by digesting pLH416 with *Hpa*I and *Sal*I, filling in and religating; it contains codons 132 to 164. Plasmid pDR617 was generated by digesting pDR609 with *Eco*RI and *Bam*HI, filling in and religating; it contains codons 47 to 76. Plasmid pTP408 was generated by digesting pLH416 with *Hpa*I and *Bam*HI, filling in and religating; it contains codons 1 to 131. Plasmid pDR618 was generated by digesting pTP408 with *Eco*RI and *Sal*I, filling in and religating; it contains codons 79 to 131. Plasmid pTP406 was generated by digesting pLH416 with *Eco*RI with *Sal*I, filling in and religating; it contains codons 79 to 164. Plasmid pTP407 was generated by digesting pLH416 with *Eco*RI and *Bam*HI, filling in and religating; it contains codons 1 to 76. Plasmid pDR643 was constructed by ligating a 126 base-pair *Hin*fI–*Hpa*I fragment containing codons 91 to 131 from pLH416 into the *Pvu*II site of pBR322.

(c) *Bacteria*

E. coli strain W3110 *lacI^q L8* (Brent & Ptashne, 1981) was used for propagation of plasmids and for growth of phage λ stocks. Strain TP302 is W3110 (*sup°*) lysogenized with P22 *Kn321 sieA44 m44* (Rennell & Poteete, 1989). Strain GM1675 (*dam-4* Δ(*lac-pro*) *thi-1 supE relA1/F′ lacI^q* Δ*M15 pro$^+$*), used for propagating phage R408 (Russel *et al.*, 1986) and generating single-stranded plasmids, was obtained from M. Marinus.

Salmonella typhimurium LT2 strains MS1362, MS1363, MS1364 and MS1365 (all *leuAam414*, bearing the amber suppressor alleles *supD*, *supE*, *supF* and *supJ*, respectively), DB7000 (*leuAam414*), MS1868 (*leuAam414 r$^-$m$^+$*) and MS2310 (MS1868 bearing the plasmid pKM101amp^R: Youderian *et al.*, 1982) were obtained from M. Susskind. Strain TP246 is MS1868 bearing 2 plasmids: pTP298, in which the *R* and *Rz* genes of phage λ are expressed under control of $P_{lac}UV5$ (D. Herrick & A. R. Poteete, unpublished results); and the *lacI*-expressing plasmid pMS421 (Grana *et al.*, 1988). Strains TP278, TP279, TP280 and TP282 are MS2310 bearing *supE*, *supF*, *supJ* and *supD*, respectively. Strains TP308 and TP309 are MS2310 bearing plasmids pDR463 and pDR464, respectively. TP308 and TP309 were maintained by continuous passage in liquid culture for 1 year in the presence of tetracycline. During this time, variants with improved ability to tolerate the presence of the suppressor plasmids came to predominate in the cultures. Derivatives of these more plasmid-tolerant variants that subsequently lost the plasmids were isolated by growth in the absence of antibiotics and screening for tetracycline and ampicillin-sensitive clones. One isolate of each, designated TP369 and TP368, respectively, was kept for further use. Strains TP371, TP372, TP374, TP375, TP377 and TP378 are TP368 bearing plasmids pFTOR1 Δ26 (ARG), pDR621, pDR622, pGFIB : sup HIS A, pGFIB : sup GLU and pGFIB : sup LYS, respectively. Strain TP376 is TP369 bearing pGFIB : sup PRO H.

The *Salmonella* strains bearing suppressor plasmids tended to lose their ability to suppress amber mutations unless handled carefully. All such strains were stored with cryoprotectant at $-80\,^{\circ}$C. Single colonies were obtained by streaking from the frozen cultures on LB agar plates and growing at 37 °C; individual colonies were then streaked on LB agar plates supplemented with antibiotics and incubated at 37 °C. The resulting colonies were then streaked on minimal agar plates containing antibiotics and incubated at 37 °C; single colonies thus obtained were then used to inoculate liquid cultures. If no colonies grew on the minimal plates within 36 h, a colony from the corresponding LB agar plate with antibiotic was streaked on minimal agar supplemented with 20 μg leucine/ml. Colonies from the minimal plates (without, or, if necessary, with leucine) were used to inoculate liquid cultures. Strains TP278, TP279, TP280, TP282, TP375, TP376 and TP377 were grown in LB supplemented with ampicillin. Strains TP308 and TP309 were grown in LB supplemented with tetracycline; TP374 was grown in LB. Strains TP371 and TP378 were grown in minimal medium supplemented with ampicillin. Strain TP372 was grown in minimal medium supplemented with tetracycline. Liquid

cultures were grown at 37 °C to densities of approximately 2×10^8/ml, tested for their ability to plate a battery of tester phages bearing amber mutations, and stored at 4 °C. Active cultures could generally be used for as long as 14 days afterwards. It was usually, but not always, possible to obtain a new active culture by 50-fold dilution of the old one in fresh medium and growth at 37 °C to 2×10^8/ml.

(d) *Determination of suppression patterns*

Phage stocks were grown by inoculating 30 ml of LB supplemented with 10 mM-sodium citrate with 1 ml of a culture of suppressor bacteria grown as described above, but to a density of approximately 5×10^8/ml, and a single plaque. (The use of citrate slightly improves the growth of the hybrid P22 *e416* derivatives, which lack gene *15* function (Casjens *et al.*, 1989).) Following aeration at 30 °C for 24 h, the culture was shaken with chloroform for an additional 30 min. Debris was removed by centrifugation at 7000 revs/min in a Sorvall SS34 rotor for 5 min, and phage were pelleted from the supernatant by centrifugation at 15,000 revs/min for 90 min in the same rotor. Phage pellets were resuspended with 2 ml of BS.

Portions (0·3 ml) of cultures of suppressor strains, grown as described above, as well as MS2310 and TP246, were mixed with 2·5 ml of molten top agar and poured on λ agar plates. The top agar was allowed to harden for 5 to 10 min at room temperature, and 7-μl portions of phage suspensions at 10^3, 10^4 and 10^5 plaque-forming units/ml were spotted on the surface. Three sets of plates were done for each phage: duplicate plates for incubation at 37 °C and single plates for incubation at 25 °C. After the spots dried, plates were incubated in an inverted position for 16 h, then scored by 2 individuals independently; scores were compared, and any discrepancies were resolved by re-examination of the plates. The plaque-forming ability of each mutant was assessed relative to that of the control wild-type hybrid phage on the same host on the same day. It was designated ++ if the plaques were about the same size as the control; + if they were significantly smaller; ± if they were of such a small size that it was difficult to discern individual plaques; and − if no plaques were produced at all. A 5th plating phenotype was occasionally seen: plaques of sufficient size to be scored as +, but having a hazy morphology that made them less visible than others of the same size; such cases were scored ±. The absolute sizes of plaques in the ++ and + categories varied from one day to another. However, the size of a mutant phage plaque as a fraction of that of the wild-type control was relatively stable. If plating conditions were changed, for instance by varying the density of the plating culture or the moisture of the plates, or the level of suppressor activity of the bacterial strain, it was possible in some cases to observe changes of one grade up or down in the score of a mutant phage.

(e) *Construction of amber mutant phages*

Amber mutations were introduced into each codon in the T4 lysozyme gene by the use of mismatched oligonucleotide primers. Primers were, in most cases, designed to carry the amber codon TAG with the outermost mismatch on either side flanked by 8 properly matched bases; 19-mers were thus used most frequently.

The targets for mutagenesis *in vitro* were, in most cases, gapped duplex DNA (Kramer & Fritz, 1987). These templates were made by mixing single-stranded, non-*dam*-methylated pLH416 with purified, large, restriction enzyme-generated double-stranded fragments of the plasmid, denaturing and reannealing. Gaps were designed to leave only the portion of the lysozyme targeted for mutagenesis in the single-stranded state. Thus, for example, in the cases of codons 81 to 117, most of the mutagenic oligonucleotides were annealed to a gapped duplex preparation in which the single-stranded part corresponded to the sequences between the *Eco*RI and *Hpa*I sites in the lysozyme gene (see Fig. 1). In 46 cases, however, the mutagenesis target was simply single-stranded, non-*dam*-methylated pLH416. In these cases, mutagenesis was carried out with the use of T7 DNA polymerase, as described by Bebenck & Kunkel (1989).

Clones bearing plasmids with amber mutations in the lysozyme gene were identified, and the mutant alleles were crossed into P22 *Kn321 sieA44 m44*, as described (Rennell & Poteete, 1989). In some cases, for instance those of codons 162 to 164, in which amber mutations do not inactivate lysozyme function even in a non-suppressing host, it was found to be advantageous to employ high-efficiency mutagenesis procedures and identify amber mutation-bearing clones directly by sequencing.

(f) *Sequencing*

The sequence of phage DNA in the vicinity of each amber mutation was determined by extension of labeled primers with reverse transcriptase in the presence of dideoxynucleotides (Inoue & Cech, 1985). DNA was purified from phage stocks as described (Rennell & Poteete, 1989); it was digested with *Hin*dIII and *Bgl*II or with *Hin*dIII alone, heat-denatured and used without further purification in the sequencing reactions. Primers used for sequencing were:

(A) 5′ CAGCAGTAACGGAATC 3′,
(B) 5′ CAAAAAGTCCATCACT 3′,
(C) 5′ TCTGAGAAATGCTAAA 3′, and
(D) 5′ CAAAAACGCTGGGATG 3′;

they were labeled by polynucleotide kinase-catalyzed phosphorylation with [γ-^{32}P]ATP, and used following extractions with phenol and ether.

Annealing mixtures contained 4 μl of template DNA (about 0·5 μg), 3 μl of primer-labeling reaction mixture (1 pmol primer), 2 μl of $5\times$AB, and 1 μl of water. They were incubated in boiling water for 3 min, quickly spun down and placed in crushed dry ice until frozen solid, then thawed at 0 °C.

Sequencing reaction mixtures contained 2 μl of annealing mixture, 1 μl of dNTP mixture, 1 μl of ddNTP mixture, and 1 μl of reverse transcriptase mixture. They were incubated at 50 °C; after 15 min, 1 μl of reverse transcriptase mixture was added and incubation was continued for an additional 15 min. The reactions were stopped by addition of 6 μl of gel loading buffer and incubation in a 90 °C water bath for 3 min. Samples of 3 μl were subjected to electrophoresis in an 8% (w/v) acrylamide/0·4% (w/v) bis-acrylamide sequencing gel.

In all cases, the presence of amber codons at the expected positions, as well as the absence of other mutations in the vicinity, was verified by sequencing. The segments of the lysozyme gene sequenced varied among the different mutants, but generally included the entire segment that was exposed as a single strand in the gapped circular duplex template. It always included the entire segment into which the conditionally lethal mutation was mapped in the phage by marker rescue tests (described below).

In 13 cases, the mutagenesis procedures resulted in the introduction of mutations in addition to or instead of the intended amber mutation. Most (9) of these appeared to be primer-directed, resulting from partial homology of the mutagenic primer to secondary sites in the lysozyme gene. In most cases, repetition of the mutagenesis procedure (often, in these cases, employing a lower concentration of primer) sufficed to generate the amber mutant without secondary mutations. In 2 cases, those of am57 and am164, the primer was redesigned as a 31-mer with the amber codon in the middle; the longer primers introduced the desired amber mutations without secondary mutations.

(g) *Marker rescue tests*

To provide assurance that the sequenced amber mutations at the intended codons were solely responsible for the plating phenotypes of the phages bearing them, marker rescue tests were carried out. These were done by spotting 10 μl portions of phage suspensions at titers of 10^6/ml on lawns of MS2310 (*sup°*), with and without plasmids, described above, containing parts of the lysozyme gene (see Fig. 1). The ability of the phage to form plaques at an elevated frequency on a plasmid-bearing host was interpreted as signifying that the phage contained no deleterious mutations outside the chromosomal segment represented in the plasmid. All of the amber mutants behaved as expected in these tests; marker rescue was detected, in general, if the amber codon was 2 codons or more away from either end of the chromosomal segment represented in the plasmid (not shown). The amber mutations located near the ends of small segments were mapped by marker rescue from plasmids bearing larger, overlapping segments.

3. Results and Discussion

A collection of 163 phages bearing amber mutations in the T4 lysozyme gene was plated on amber suppressor strains and scored for plaque-forming ability at 37 °C and 25 °C. The results are shown in Table 1; plating phenotypes are illustrated in Figure 2. Plaque-forming ability was judged as described in Materials and Methods. In what follows, scores of ++ or + will be considered positive, and $\pm$ or $-$ will be considered negative; the descriptions of phenotypes refer to plating at 37 °C, unless otherwise specified.

(a) *Interpretation of amber mutant suppression patterns*

In principle, the experiment summarized in Table 1 tests the qualitative functionality, relative to wild-type, of 2015 mutant lysozymes bearing single

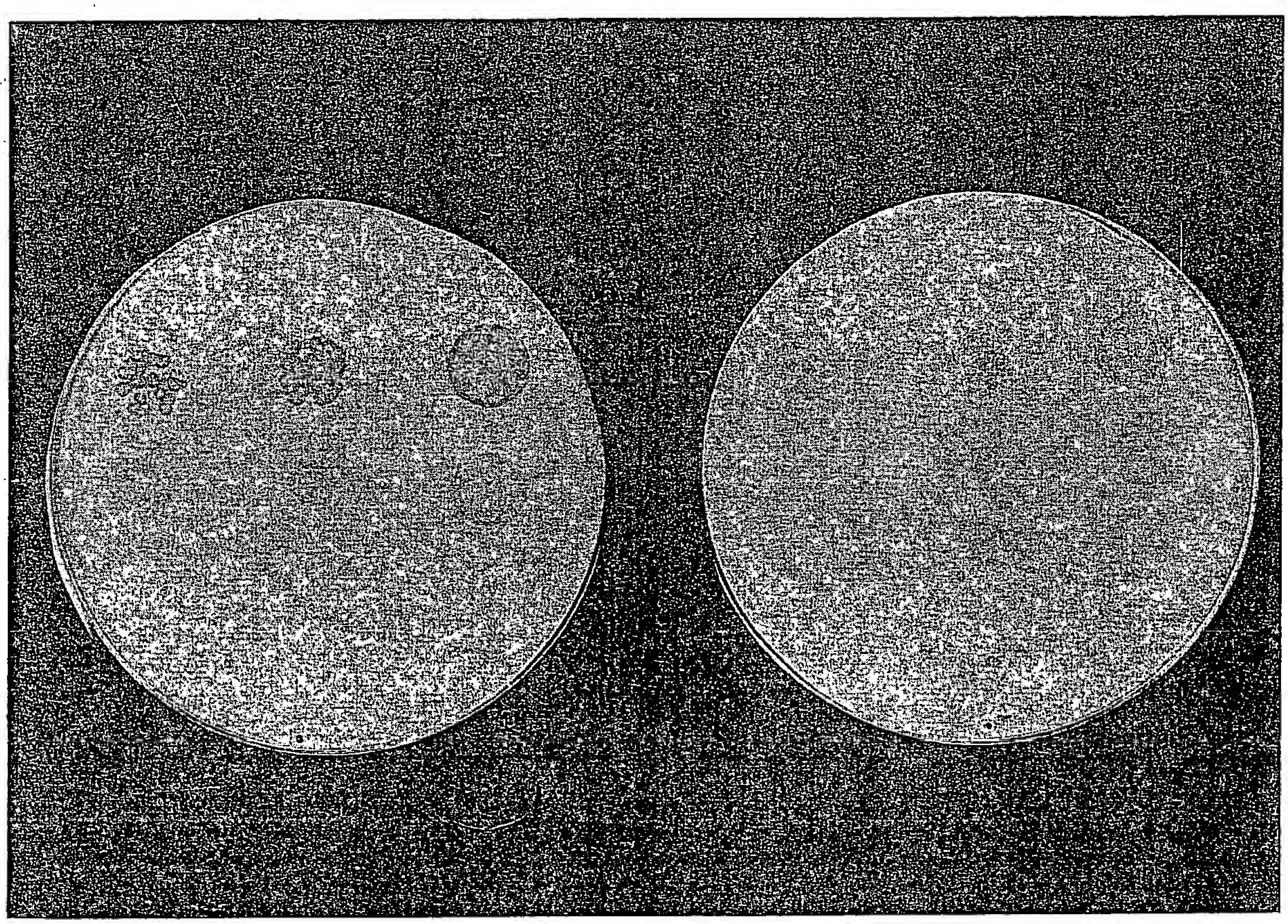


Figure 2. Plating phenotypes of derivatives of P22 *e416* bearing amber mutations in gene *e*. In the plates shown, phages bearing amber mutations in codons 2, 59, 161 and 11 (in order from top to bottom) were plated on strains TP280 (Leu-inserting; left side) and TP246 (fully permissive, expresses R and R_Z genes of phage λ; right side) at 37 °C. Phage suspensions had titers of 10^3, 10^4 and 10^5/ml (in order from left to right). The results illustrate the scores ++, +, $\pm$ and $-$, respectively.

Table 1
Suppression patterns of T4 lysozyme amber mutants

	Su-	Gly	Ala	Leu	Gln	Cys	Ser	Tyr	Phe	His	Pro	Glu	Arg	Lys	
416	++	++	++	++	++	++	++	++	++	++	++	++	++	++	37°C
WT	++	++	++	++	++	++	++	++	++	++	++	++	++	++	25°C
2	−	++	++	++	++	+	++	++	++	++	++	++	++	+	
Asn	−	++	++	++	++	±	++	++	+	++	++	++	++	+	
3	−	++	++	++	++	+	++	++	++	++	++	+	+	±	
Ile	−	++	+	++	++	±	++	++	+	++	+	+	+	±	
4	−	++	++	++	++	++	++	++	++	++	++	++	++	++	
Phe	−	++	++	++	++	+	++	++	++	++	++	++	++	++	
5	−	++	++	++	++	++	++	++	++	++	++	++	++	++	
Glu	±	++	++	++	++	+	++	++	++	++	++	++	++	++	
6	−	+	++	++	++	+	++	+	+	+	+	−	−	−	
Met	±	++	++	++	++	+	++	++	++	++	++	+	−	+	
7	−	++	++	++	++	±	++	++	++	+	−	−	−	−	
Leu	−	++	++	++	++	±	++	++	+	+	±	+	−	±	
8	+	++	++	++	++	++	++	++	++	++	++	++	++	++	
Arg	+	++	++	++	++	++	++	++	++	++	++	++	++	++	
9	±	++	++	++	++	++	++	++	++	++	++	++	++	++	
Ile	+	++	++	++	++	++	++	++	++	++	++	++	++	++	
10	−	−	−	+	+	−	±	−	−	−	−	++	−	−	
Asp	−	+	+	+	+	−	++	−	−	+	−	++	−	−	
11	−	−	−	−	−	−	−	−	−	−	−	++	−	−	
Glu	−	−	−	−	−	−	−	−	−	−	−	++	−	−	
12	−	++	++	++	++	+	++	+	++	++	−	++	++	±	
Gly	−	++	++	++	++	+	++	+	++	++	−	++	++	+	
13	−	++	++	++	++	++	++	++	++	++	++	++	++	++	
Leu	−	++	++	++	++	+	++	++	++	++	++	++	++	++	
14	−	++	++	++	++	++	++	++	++	++	++	++	++	++	
Arg	−	++	++	++	++	+	++	++	++	++	+	++	++	++	
15	−	++	++	++	++	++	++	++	++	++	++	++	++	+	
Leu	−	++	++	++	++	+	++	++	++	++	++	++	++	+	
16	−	++	++	++	++	++	++	++	++	++	++	++	++	++	
Lys	−	++	++	++	++	+	++	++	++	++	++	++	++	++	
17	−	+	++	++	++	±	+	++	++	++	++	±	±	−	
Ile	−	++	++	++	++	±	++	++	+	++	++	+	±	±	
18	−	+	++	++	++	±	+	++	++	++	±	+	+	−	
Tyr	−	±	++	+	+	±	+	++	++	++	±	+	+	±	
19	±	++	++	++	++	++	++	++	++	++	++	++	++	++	
Lys	+	++	++	++	++	++	++	++	++	++	++	++	++	++	
20	−	+	++	−	±	++	±	±	−	±	±	++	−	−	
Asp	−	±	++	−	±	+	−	−	−	±	+	+	−	−	
21	±	++	++	++	++	++	++	++	++	++	++	++	++	++	
Thr	±	++	++	++	++	+	++	++	++	++	++	++	++	++	
22	±	++	++	++	++	++	++	++	++	++	++	++	++	++	
Glu	±	++	++	++	++	+	++	++	++	++	++	++	++	++	
23	−	++	++	++	++	+	++	++	++	++	+	++	++	+	
Gly	−	++	++	++	++	±	++	++	++	++	+	++	++	+	
24	−	++	++	++	++	+	++	++	++	++	+	+	++	+	
Tyr	−	++	++	++	++	±	++	++	++	++	+	+	++	+	

Table 1 *(continued)*

	Su-	Gly	Ala	Leu	Gln	Cys	Ser	Tyr	Phe	His	Pro	Glu	Arg	Lys
25	−	++	++	++	++	++	++	++	++	++	++	++	++	++
Tyr	±	++	++	++	++	+	++	++	++	++	++	++	++	++
26	−	++	++	−	−	±	++	±	±	+	±	−	−	−
Thr	−	++	++	−	−	±	++	−	−	+	±	−	−	−
27	−	+	++	+	+	++	++	+	+	±	−	±	−	−
Ile	−	++	++	++	++	+	++	++	++	++	−	+	−	−
28	−	++	++	−	−	±	++	−	−	−	−	−	−	−
Gly	−	++	++	−	−	±	++	±	±	+	−	−	−	−
29	−	++	++	++	++	++	++	++	++	++	++	++	+	±
Ile	−	++	++	++	++	+	++	++	++	++	+	++	+	±
30	−	++	++	−	−	−	±	−	−	−	−	−	−	−
Gly	−	++	++	−	−	−	±	−	−	−	−	−	−	−
31	−	+	++	++	++	±	++	++	+	++	−	±	++	±
His	−	++	++	++	++	±	++	+	+	++	−	+	++	±
32	+	++	++	++	++	++	++	++	++	++	++	++	++	++
Leu	+	++	++	++	++	++	++	++	++	++	++	++	++	++
33	−	++	++	++	++	++	++	++	++	++	++	+	±	±
Leu	−	++	++	++	++	+	++	+	++	++	++	++	+	±
34	−	++	++	++	++	++	++	++	++	++	++	++	++	+
Thr	−	++	++	++	++	+	++	++	++	++	++	++	++	+
35	−	++	++	++	++	++	++	++	++	++	++	++	++	++
Lys	−	++	++	++	++	+	++	++	++	++	++	++	++	++
36	−	++	++	++	++	++	++	++	++	++	++	++	++	+
Ser	−	++	++	++	++	+	++	++	++	++	++	++	++	+
37	±	++	++	++	++	++	++	++	++	++	++	++	++	+
Pro	±	++	++	++	++	+	++	++	++	++	++	++	++	+
38	−	++	++	++	++	++	++	++	++	++	++	++	++	+
Ser	−	++	++	++	++	+	++	++	++	++	++	++	++	+
39	−	++	++	++	++	++	++	++	++	++	++	++	++	+
Leu	−	++	++	++	++	+	++	++	++	++	++	++	++	+
40	±	++	++	++	++	++	++	++	++	++	++	++	++	++
Asn	−	++	++	++	++	+	++	++	++	++	++	++	++	++
41	+	++	++	++	++	++	++	++	++	++	++	++	++	++
Ala	+	++	++	++	++	+	++	++	++	++	++	++	++	++
42	−	++	++	++	++	++	++	++	++	++	++	+	±	±
Ala	−	++	++	++	++	+	++	++	++	++	++	++	++	++
43	−	++	++	++	++	++	++	++	++	++	++	++	++	+
Lys	−	++	++	++	++	+	++	++	++	++	++	++	++	++
44	−	++	++	++	++	++	++	++	++	++	++	++	++	+
Ser	−	++	++	++	++	+	++	++	++	++	++	++	++	++
45	±	++	++	++	++	+	++	++	++	++	+	++	++	+
Glu	−	++	++	++	++	+	++	++	++	++	++	+	++	±
46	−	+	++	+	+	+	+	±	+	±	±	−	±	−
Leu	−	++	++	++	++	+	++	+	++	++	+	+	++	±
47	−	++	++	++	++	++	++	++	++	++	+	++	++	++
Asp	−	++	++	++	++	++	++	++	++	++	++	++	++	+
48	+	++	++	++	++	++	++	++	++	++	++	++	++	++
Lys	+	++	++	++	++	++	++	++	++	++	++	++	++	++

Table 1 *(continued)*

	Su-	Gly	Ala	Leu	Gln	Cys	Ser	Tyr	Phe	His	Pro	Glu	Arg	Lys
49	−	++	++	++	++	++	++	++	++	++	++	++	++	++
Ala	−	++	++	++	++	+	++	++	++	++	++	++	++	++
50	±	++	++	++	++	++	++	++	++	++	±	++	++	+
Ile	+	++	++	++	++	+	++	++	++	++	+	++	++	++
51	−	++	+	++	+	±	++	+	+	+	+	+	++	+
Gly	−	++	++	++	++	±	++	++	++	++	+	+	++	±
52	−	++	++	++	++	++	++	++	++	++	++	++	++	++
Arg	−	++	++	++	++	+	++	++	++	++	++	++	++	+
53	−	++	+	++	+	±	++	++	+	+	+	+	+	±
Asn	−	++	+	++	++	±	++	++	+	++	+	±	+	±
54	−	++	++	++	++	++	++	++	++	++	++	+	++	++
Cys	−	++	++	++	++	+	++	++	++	++	++	++	++	++
55	−	++	++	++	++	++	++	++	++	++	++	++	++	++
Asn	−	++	++	++	++	+	++	++	++	++	++	++	++	++
56	−	++	++	++	++	++	++	++	++	++	+	++	++	++
Gly	−	++	++	++	++	+	++	++	++	++	++	++	++	++
57	−	++	++	++	++	++	++	++	++	++	++	++	++	++
Val	−	++	++	++	++	+	++	++	++	++	++	++	++	++
58	−	±	++	±	±	+	+	±	++	±	+	−	−	±
Ile	−	++	++	++	++	++	++	++	++	++	++	+	±	±
59	−	++	+	+	+	±	++	+	±	++	+	+	++	±
Thr	±	++	++	++	++	+	++	++	++	++	++	++	++	++
60	±	++	++	++	++	++	++	++	++	++	++	++	++	++
Lys	±	++	++	++	++	+	++	++	++	++	++	++	++	++
61	−	++	++	++	++	+	++	++	++	++	++	++	++	+
Asp	−	++	++	++	++	±	++	++	++	++	++	++	++	+
62	±	++	++	++	++	++	++	++	++	++	±	++	±	±
Glu	+	++	++	++	++	+	++	++	++	++	+	++	++	++
63	−	++	++	++	++	++	++	++	++	+	−	±	±	±
Ala	−	++	++	++	++	+	++	++	++	++	+	++	++	+
64	−	++	++	++	++	++	++	++	++	++	+	++	++	++
Glu	−	++	++	++	++	+	++	++	++	++	++	++	++	++
65	±	++	++	++	++	++	++	++	++	++	++	++	++	++
Lys	±	++	++	++	++	+	++	++	++	++	++	++	++	++
66	−	−	+	++	++	±	±	++	++	+	−	−	+	±
Leu	−	++	++	++	++	±	++	++	++	++	±	++	++	+
67	−	++	++	++	++	±	++	++	++	++	−	+	+	±
Phe	−	++	++	++	++	±	++	++	++	++	−	++	++	±
68	±	++	++	++	++	++	++	++	++	++	++	++	++	++
Asn	+	++	++	++	++	+	++	++	++	++	++	++	++	++
69	±	++	++	++	++	++	++	+	++	++	++	++	++	++
Gln	±	++	++	++	++	+	++	++	++	++	++	++	++	++
70	−	++	++	++	++	±	++	+	±	++	±	++	±	±
Asp	−	++	++	++	++	±	++	++	±	++	±	++	±	±
71	±	++	++	++	++	++	++	++	++	++	+	++	++	++
Val	+	++	++	++	++	+	++	++	++	++	+	++	++	++
72	±	++	++	++	++	++	++	+	+	++	++	++	++	++
Asp	+	++	++	++	++	+	++	++	++	++	++	++	++	++

Table 1 *(continued)*

	Su-	Gly	Ala	Leu	Gln	Cys	Ser	Tyr	Phe	His	Pro	Glu	Arg	Lys
73	+	++	++	++	++	++	++	+	++	++	++	++	++	++
Ala	+	++	++	++	++	+	++	++	++	++	++	++	++	++
74	−	++	++	++	++	++	++	++	++	++	+	+	++	++
Ala	−	++	++	++	++	+	++	++	++	++	+	+	++	++
75	−	++	++	++	++	+	++	++	+	++	±	+	++	++
Val	−	++	++	++	++	±	++	++	++	++	±	+	++	++
76	−	++	++	++	++	++	++	++	++	++	++	++	++	++
Arg	−	++	++	++	++	++	++	++	++	++	++	++	++	++
77	−	++	++	++	++	++	++	++	++	++	++	++	++	++
Gly	−	++	++	++	++	+	++	++	++	++	++	+	++	++
78	−	+	++	+	+	+	++	±	+	±	+	−	−	±
Ile	−	++	++	++	++	+	++	++	++	+	++	+	±	±
79	−	++	++	++	++	+	++	++	+	++	++	++	++	++
Leu	−	++	++	++	++	+	++	++	++	++	++	++	++	++
80	−	++	++	++	++	+	++	++	+	++	++	++	++	++
Arg	−	++	++	++	++	+	++	++	++	++	++	++	++	++
81	±	++	++	++	++	+	++	++	+	++	+	++	++	++
Asn	±	++	++	++	++	+	++	++	++	++	+	++	++	++
82	−	++	++	++	++	+	++	++	+	++	++	++	++	++
Ala	−	++	++	++	++	+	++	++	++	++	++	++	++	++
83	±	++	++	++	++	++	++	++	++	++	++	++	++	++
Lys	±	++	++	++	++	+	++	++	++	++	++	++	++	++
84	−	±	++	+	+	++	+	+	++	+	−	−	−	−
Leu	−	++	++	++	++	+	++	++	++	++	−	+	±	±
85	−	++	++	++	++	++	++	++	++	++	+	++	++	++
Lys	−	++	++	++	++	+	++	++	++	++	++	++	++	++
86	−	++	++	++	++	++	++	++	++	++	++	++	++	++
Pro	−	++	++	++	++	+	++	++	++	++	++	++	++	++
87	−	++	++	++	+	±	++	++	+	++	++	−	±	±
Val	−	++	++	++	++	±	++	++	+	++	+	+	+	±
88	−	++	++	++	++	++	+	++	++	++	−	+	++	−
Tyr	−	++	++	++	++	+	++	++	++	++	±	++	++	±
89	−	++	+	++	++	+	+	++	+	++	++	+	++	+
Asp	−	++	++	++	++	±	++	++	+	++	++	+	++	±
90	±	++	++	++	++	++	++	++	++	++	++	++	++	++
Ser	±	++	++	++	++	++	++	++	++	++	++	++	++	++
91	−	++	++	++	++	++	+	+	++	++	−	±	−	±
Leu	−	++	++	++	++	+	++	++	++	++	++	++	++	±
92	−	++	++	++	++	++	++	++	++	++	++	++	++	±
Asp	−	++	++	++	++	+	++	++	++	++	++	++	+	±
93	±	++	++	++	++	++	++	++	++	++	++	++	++	++
Ala	−	++	++	++	++	++	++	++	++	++	++	++	++	++
94	−	++	+	++	++	±	++	++	+	+	++	+	+	+
Val	−	++	++	++	++	±	++	++	+	+	++	+	+	+
95	−	++	+	++	++	−	+	−	−	±	±	±	++	+
Arg	−	++	++	++	++	±	++	±	−	+	++	+	++	+
96	−	++	+	++	++	±	++	++	+	+	±	+	+	+
Arg	−	++	++	++	++	±	++	++	+	+	±	±	+	+

Table 1 *(continued)*

	Su-	Gly	Ala	Leu	Gln	Cys	Ser	Tyr	Phe	His	Pro	Glu	Arg	Lys
97	−	++	++	++	++	+	++	++	++	+	++	++	++	+
Cys	−	++	++	++	++	±	++	++	++	+	++	++	++	+
98	−	++	+	±	±	+	++	−	−	−	+	−	−	±
Ala	−	++	++	+	++	±	++	−	−	−	+	−	−	±
99	−	±	++	++	++	+	+	++	++	++	−	+	−	+
Leu	−	++	++	++	++	+	++	++	++	++	−	+	−	+
100	−	++	++	++	++	+	++	++	++	++	±	++	+	++
Ile	−	++	++	++	++	+	++	++	++	++	−	++	++	+
101	−	++	++	+	+	+	++	+	+	+	−	±	−	±
Asn	−	++	++	++	++	+	++	±	++	+	−	++	±	+
102	−	+	++	++	++	+	++	−	+	+	−	−	−	±
Met	−	++	++	++	++	+	++	−	++	++	+	−	−	±
103	−	++	++	++	++	+	++	++	+	++	−	+	+	±
Val	−	++	++	++	++	±	++	+	+	++	−	++	+	±
104	−	+	++	+	+	+	+	++	++	+	−	±	−	±
Phe	−	+	+	+	±	+	±	++	+	+	−	±	−	−
105	−	++	++	++	++	+	++	−	−	++	−	++	−	±
Gln	−	+	++	++	++	±	++	−	−	+	−	+	−	±
106	−	++	++	++	++	+	++	±	++	++	−	+	±	±
Met	−	++	++	++	++	+	++	−	++	+	−	+	−	+
107	−	++	++	+	+	+	++	++	±	+	±	+	+	±
Gly	−	++	++	+	+	±	++	+	+	++	±	+	±	±
108	−	++	++	++	++	++	++	++	++	++	++	++	++	++
Glu	−	++	++	++	++	+	++	++	++	++	++	++	++	+
109	±	++	++	++	++	++	++	++	++	++	++	++	++	++
Thr	±	++	++	++	++	++	++	++	++	++	++	++	++	++
110	−	++	++	++	++	++	++	+	+	++	++	++	++	++
Gly	−	++	+	++	++	+	++	+	+	+	++	++	++	++
111	−	++	++	+	+	+	++	+	+	−	++	±	−	±
Val	−	++	++	++	++	+	++	+	++	−	++	+	−	+
112	−	++	++	++	++	+	++	++	++	++	++	++	++	+
Ala	−	++	++	++	++	±	++	++	++	++	++	++	+	±
113	−	++	++	++	++	+	++	++	++	++	++	++	++	+
Gly	−	++	++	++	++	±	++	++	++	++	++	++	+	±
114	−	±	+	+	+	+	+	++	++	+	−	−	±	−
Phe	−	±	±	+	+	+	±	++	++	±	−	−	−	−
115	−	++	++	++	++	++	++	++	++	++	++	++	++	++
Thr	−	++	++	++	++	+	++	++	++	++	++	++	++	+
116	±	++	++	++	++	++	++	++	++	++	++	++	++	++
Asn	±	++	++	++	++	++	++	++	++	++	++	++	++	+
117	−	++	++	+	+	±	++	±	±	±	±	±	−	±
Ser	−	++	++	±	±	−	++	−	±	−	±	−	−	±
118	−	++	++	++	++	++	++	++	+	++	±	±	+	±
Leu	−	++	++	++	++	+	++	+	+	++	+	+	++	±
119	−	++	++	++	++	++	++	++	+	++	++	++	++	++
Arg	−	++	++	++	++	+	++	++	++	++	++	++	++	++
120	±	++	++	++	++	++	++	++	+	++	++	++	++	++
Met	−	++	++	++	++	+	++	++	++	++	++	++	++	++

Table 1 *(continued)*

	Su-	Gly	Ala	Leu	Gln	Cys	Ser	Tyr	Phe	His	Pro	Glu	Arg	Lys
121	−	++	++	+	+	+	++	+	+	±	−	−	−	±
Leu	−	++	++	++	++	±	++	++	+	+	±	+	−	+
122	−	++	++	++	++	+	++	++	++	++	++	+	++	+
Gln	−	++	++	++	++	±	++	++	+	++	+	+	++	+
123	−	++	++	++	++	++	++	++	++	++	+	++	++	++
Gln	−	++	++	++	++	±	++	++	++	++	++	++	++	++
124	−	++	++	++	++	+	++	++	++	++	+	+	++	+
Lys	−	++	++	++	++	±	++	++	++	++	+	+	++	+
125	−	++	++	++	++	+	++	++	++	++	±	+	+	+
Arg	−	++	++	++	++	±	++	++	+	++	±	+	+	+
126	−	+	+	++	+	±	+	++	++	++	+	+	±	−
Trp	−	++	++	++	++	−	++	++	++	++	+	+	++	±
127	−	++	++	++	++	+	++	++	++	++	++	++	++	+
Asp	−	++	++	++	++	±	++	++	+	++	++	++	++	±
128	±	++	++	++	++	++	++	++	++	++	++	++	++	++
Glu	±	++	++	++	++	+	++	++	++	++	++	++	++	++
129	−	++	++	+	+	+	++	+	++	±	±	±	−	±
Ala	−	++	++	++	++	±	++	+	+	++	+	+	−	±
130	−	++	++	++	++	++	++	++	++	++	++	++	++	++
Ala	−	++	++	++	++	+	++	++	++	++	++	++	++	++
131	−	++	++	++	++	++	++	++	++	++	++	++	++	++
Val	−	++	++	++	++	+	++	++	++	++	++	++	++	++
132	−	++	++	++	++	+	++	+	+	++	+	++	+	+
Asn	−	++	++	++	++	±	++	++	++	++	+	++	+	+
133	−	±	++	+	+	+	+	++	++	−	−	−	−	±
Leu	−	++	++	++	++	+	++	++	++	++	±	+	−	++
134	−	++	++	++	++	++	++	++	++	++	++	++	++	++
Ala	−	++	++	++	++	+	++	++	++	++	++	++	++	++
135	−	++	++	++	++	++	++	++	++	++	++	++	++	++
Lys	−	++	++	++	++	+	++	++	++	++	++	++	++	++
136	−	++	++	+	±	−	++	−	−	−	−	−	±	±
Ser	−	++	++	±	+	−	++	−	−	−	−	−	−	+
137	−	++	+	++	+	±	++	+	+	++	±	±	++	+
Arg	−	++	+	++	++	±	++	++	+	++	±	−	+	+
138	−	±	±	+	−	−	−	+	+	−	−	−	−	−
Trp	−	−	±	+	−	−	−	++	±	−	−	−	−	−
139	−	++	++	++	++	+	++	++	++	++	++	++	++	++
Tyr	−	++	++	++	++	±	++	++	+	++	+	+	++	++
140	−	++	+	++	++	+	++	++	+	++	++	+	++	++
Asn	−	++	++	++	++	±	++	++	+	++	+	±	++	++
141	±	++	++	++	++	++	++	++	++	++	+	++	++	++
Gln	±	++	++	++	++	+	++	++	+	++	+	++	++	++
142	−	++	++	+	+	+	++	+	+	±	−	+	−	±
Thr	−	++	++	+	+	±	++	±	+	−	±	±	−	±
143	−	++	++	++	++	++	++	++	++	++	++	++	++	++
Pro	−	++	++	++	++	+	++	++	++	++	++	++	++	++
144	−	++	++	++	++	+	++	++	++	+	+	+	+	+
Asn	−	++	++	++	++	±	++	++	++	++	++	+	+	+

Table 1 *(continued)*

	Su-	Gly	Ala	Leu	Gln	Cys	Ser	Tyr	Phe	His	Pro	Glu	Arg	Lys
145	−	±	+	+	±	±	+	−	±	++	+	±	++	+
Arg	±	±	+	±	±	+	+	−	±	+	+	±	++	+
146	−	++	++	+	+	+	++	±	+	±	++	−	−	−
Ala	−	++	++	±	+	±	++	−	±	±	++	±	−	−
147	−	++	++	++	++	+	++	++	++	++	++	+	++	+
Lys	−	++	++	++	++	±	++	++	+	++	+	±	++	+
148	−	+	++	++	+	±	++	±	±	++	+	−	++	+
Arg	−	++	++	++	++	±	++	+	+	++	+	−	++	+
149	−	−	++	+	+	±	±	−	−	−	−	−	−	−
Val	−	++	++	++	++	−	++	−	−	−	±	−	−	−
150	−	++	++	++	++	+	++	++	++	++	−	++	+	+
Ile	−	++	++	++	++	±	++	++	++	++	−	++	++	+
151	±	++	++	++	++	++	++	++	++	++	++	++	++	++
Thr	±	++	++	++	++	+	++	++	++	++	++	++	++	++
152	−	++	++	++	++	+	++	−	−	+	−	−	−	±
Thr	−	++	++	++	++	±	++	−	±	++	−	−	−	±
153	−	+	++	++	+	+	+	++	++	++	−	−	−	±
Phe	−	++	++	++	++	±	++	++	++	++	−	−	−	±
154	±	++	++	++	++	++	++	++	++	++	±	++	++	++
Arg	−	++	++	++	++	+	++	++	++	++	−	++	++	++
155	−	++	++	++	++	++	++	++	++	++	±	++	++	++
Thr	−	++	++	++	++	+	++	++	++	++	−	++	++	+
156	−	++	++	+	+	+	++	−	−	−	−	±	+	±
Gly	−	++	++	++	++	+	++	−	−	+	+	++	++	+
157	−	++	+	++	++	+	++	++	++	+	−	+	++	+
Thr	−	++	+	++	++	±	++	++	+	+	−	+	+	+
158	−	++	++	++	++	++	++	++	++	++	++	++	++	++
Trp	±	++	++	++	++	++	++	++	++	++	++	++	++	+
159	±	++	++	++	++	++	++	++	++	++	+	++	++	++
Asp	+	++	++	++	++	++	++	++	++	++	++	++	++	++
160	−	++	++	++	++	+	++	++	+	+	+	+	+	±
Ala	+	++	++	++	++	+	++	++	+	++	+	+	++	++
161	−	−	−	±	−	−	−	++	+	±	−	−	−	−
Tyr	±	+	±	+	±	±	±	++	+	++	±	±	+	+
162	++	++	++	++	++	++	++	++	++	++	++	++	++	++
Lys	++	++	++	++	++	++	++	++	++	++	++	++	++	++
163	++	++	++	++	++	++	++	++	++	++	++	++	++	++
Asn	++	++	++	++	++	++	++	++	++	++	++	++	++	++
164	++	++	++	++	++	++	++	++	++	++	++	++	++	++
Leu	++	++	++	++	++	++	++	++	++	++	++	++	++	++

Plaque-forming ability was determined as described in Materials and Methods. Top rows refer to plating at 37 °C, bottom rows to plating at 25 °C. WT, wild-type.

amino acid substitutions. However, interpretation of amber mutant suppression patterns is not completely straightforward, for a number of reasons.

First, the efficiency of amber suppressors is not 100%, and varies from one suppressor to another. Thus, the phenotype of an amber mutant growing on an amber suppressor strain results from the combined effects of alteration of the lysozyme and under-expression relative to the wild-type. On the other hand, the suppressors used in these studies have been characterized as relatively efficient (Winston *et al.*, 1979; McClain & Foss, 1988; Kleina *et al.*, 1990). Moreover, studies with hybrid phages engineered to produce varying amounts of wild-type lysozyme indicate that, for a mutation to be scored as defective, it must reduce total lysozyme activity to less than 3% of the amount produced by the wild-type hybrid phage (Knight *et al.*, 1987; our unpublished results). Finally, as indicated in Table 1, every suppressor works with mutants bearing amber codons at a large number of positions, including all cases in which the resultant polypeptide is supposed to have the wild-type amino acid sequence. (There are 104 cases of non-substitution in the Table.)

Second, the efficiency of amber suppressors is subject to variation from one amber codon to another. Such "context effects", though, are relatively small (generally less than 3-fold: Miller & Albertini, 1983; Bossi, 1983). Indeed, every amber mutant forms large plaques at high efficiency on at least one amber suppressor. Still, general suppressor efficiency and context effects may play a significant role in determining some of the phenotypes shown. Such a role is suggested by the observation that in 11 cases (codons 43, 46, 84, 96, 97, 98, 121, 124, 125, 133 and 147) an amber mutant phage makes a smaller than wild-type plaque on an amber suppressor strain that inserts the wild-type amino acid.

A third factor complicating interpretation of the amber mutant suppression patterns shown in Table 1 is the question of suppressor specificity. Although all of the suppressors used here have been well-characterized, two types of uncertainty remain. (1) Only a small amount of residual lysozyme activity is needed for plaque formation; therefore, an amber suppressor that inserted a particular "incorrect" amino acid only a small fraction of the time could give false positives. Indeed, the Glu-inserting suppressor has been found to insert Gln 17% of the time (Normanly *et al.*, 1990). (It should be noted, however, that the data in Table 1 indicate a number of mutants that form large plaques on the Gln-inserting suppressor strain, and no plaques at all on the Glu-inserting suppressor strain.) (2) The synthetic suppressors were characterized in *E. coli*, but tested with T4 lysozyme amber mutants in *S. typhimurium*. While we suppose that their specificities would be the same in these two closely related species, this has not been determined directly.

On the other hand, a number of observations suggest that the suppressors behave largely as expected. The mutants that plate on the fewest suppressor strains tend to plate only on those that make conservative substitutions (as well as non-substitutions of course). The exemplar of this tendency is am11. In T4 lysozyme, Glu11 is thought to be the key catalytic residue (Anderson *et al.*, 1981); as one might therefore expect, am11 plates only on the Glu-inserting suppressor. In another example, Table 1 indicates that Gly30 can be replaced only by Ala. Similarly, Tyr161 can be replaced only by Phe, and Trp138 can be replaced only with Tyr, Phe or Leu. With 13 different suppressors, and all 20 amino acids represented in the lysozyme polypeptide, 247 different kinds of substitution can be made. For any given amino acid, it is possible, by analysis of the data in Table 1, to determine how well its positions in the protein can be substituted by any of the 13 nominally represented in the collection of suppressors. Upon carrying out such an analysis (not shown), we found, for example, that alanine residues can be replaced most effectively by Gly and Ser. Similarly, serine residues can be replaced most effectively by Ala and Gly, and glycine residues by Ala and Ser. These three amino acid residues thus constitute a self-contained "exchange group", clearly related chemically by virtue of being the three smallest. Likewise, Phe and Tyr were found to constitute a self-contained exchange group. Little other useful information was obtained from such an analysis of the entire suppression pattern; it was far more informative when these relationships were examined within two groups of lysozyme residues, buried and solvent-exposed, identified by examination of the structure (see below).

A fourth complication characteristic of amber mutations is "leakiness". Bossi (1983) found that amber codons in a *lacI-Z* fusion gene gave rise to as much as 2% of the wild-type level of activity in a non-suppressing *Salmonella* background. Amber codons at eight positions in T4 lysozyme (8, 32, 41, 48, 73, 162, 163 and 164) fail to prevent plaque formation on the non-suppressor strain; phages bearing the last three of these plate particularly well. Two likely explanations for this phenomenon are: (1) the codons in question are translated through at relatively high frequency and represent positions that are tolerant of substitutions; (2) the amber mutants produce active lysozyme fragments. We believe that the first explanation pertains in the first five cases, the second in the last three. The fact that tight negative phenotypes are observed for amber mutations of most codons through 161 suggests that shorter N-terminal fragments of T4 lysozyme are not, in general, active. We tested the possibility that particular amber fragments might be active by constructing variants of pLH416 bearing amber codons at positions 71, 72 and 73, in combination with a deletion that eliminates sequences coding for residues 79 through 164. These plasmids were all able to recombine with the defective prophage P22 *Kn321* to yield phages that could form plaques on a host (TP246) that supplies lysis

functions to the infecting phage; none of these recombinants was able to form plaques on a normal host (not shown). In addition, we have found the double amber mutant P22 *e416 am73 am161* does not form plaques on a non-suppressor host (not shown); if the fragment produced by the amber codon at position 73 were active, the presence of a second amber mutation downstream would presumably make little difference. Phipps *et al.* (1987) have described an active fusion protein consisting of residues 1 to 78 of T4 lysozyme followed by 19 other amino acid residues derived from a cloning vector. The activity of such a fusion protein would appear to be difficult to reconcile with our finding that a number of residues past 78 are critical for lysozyme function. Possibly, the amino terminus of T4 lysozyme contains all the necessary residues for catalysis, but requires additional structure at the carboxy terminus for stability, and the 19 amino acid residues in the fusion protein fortuitously provide such stabilizing structure.

(b) *Overall tolerance to substitutions*

The data in Table 1 show the overall functional effects of 2015 substitutions at 163 out of the 164 positions in the T4 lysozyme polypeptide. Only 328 of these substitutions, affecting 74 of the residues, were scored as deleterious. From these results it would appear that most (89 out of 163, or 55%) of the positions in the molecule are insensitive to substitution, being able to tolerate a minimum of 13 different amino acid residues.

Tolerance to amino acid substitutions is a salient feature of proteins. That there are many sequences that can form the highly conserved globin structure became clear from the work of Perutz *et al.* (1965). Subsequent studies of over 300 mutant human hemoglobins, some defective, but most fully functional, lead to the same general conclusion (for a review, see Weatherall & Clegg, 1976). Miller and co-workers (Miller *et al.*, 1979; Kleina & Miller, 1990) pioneered the approach, used in this study, of using nonsense mutations and suppressors to test the effects of amino acid substitutions. This approach permits the study of mutant variants with no requirement that they be able to arise or survive in natural populations. By testing nonsense mutations in 141 of the 360 positions of the *E. coli lac* repressor, these investigators were able to examine the effects of 1634 single amino acid substitutions. They found that approximately 70% of the substitutions in the DNA-binding amino-terminal domain, and 30% of the others (approx. 45% overall) led to loss of function. Loeb *et al.* (1989) found that most positions in HIV-1 protease could tolerate amino acid substitutions. Sauer and co-workers (Reidhaar-Olson & Sauer, 1988; Bowie *et al.*, 1990) have employed a more radical form of mutational analysis, called combinatorial cassette mutagenesis, to study the question of tolerance to amino acid substitution. In this method, a segment of DNA sequence encoding a part of a protein is effectively randomized *in vitro*; and variants encoding functional proteins are subsequently selected and sequenced. Using such methods, these investigators have found that most residues of the phage λ repressor amino-terminal domain tolerate many substitutions.

Among the studies cited above, the results of Kleina & Miller (1990) are most precisely comparable to those reported here. Out of 132 positions represented in the collection of *lacI* amber mutations, 30 were completely insensitive to substitution; an additional 22 showed only partial loss of function. If the 37% of the residues of *lac* repressor represented in the collection of nonsense mutations are typical, then these figures can be compared to our finding that 89 out of 163 residues in T4 lysozyme can be successfully substituted by any member of the same set of 13 amino acids.

The fraction of substitutions in T4 lysozyme that are functional is a quantity that can be varied almost arbitrarily in the system described here. In constructing the hybrid phage used for these studies, we found that the level of expression of the wild-type lysozyme gene could be varied, by genetic engineering, over a nearly 1000-fold range without loss of plaque-forming ability (Knight *et al.*, 1987; Hardy & Poteete, 1991). A level near the middle of this range, comparable to the amount of P22 lysozyme made by wild-type P22, was chosen. It should be noted that, in a hybrid phage that makes less lysozyme, more substitutions are scored as deleterious; in one that makes more lysozyme, fewer substitutions are apparently deleterious (not shown).

(c) *Buried residues are sensitive to substitution*

In Figure 3, positions in T4 lysozyme that are sensitive to substitution are indicated, as well as positions occupied by residues with solvent-inaccessible side-chains. The striking correlation of these properties indicates that interior positions are, in general, more sensitive to substitution than surface residues. Two conspicuous exceptions to this generalization are Glu11 and Asp20, both of which are thought to be directly involved in catalysis (Anderson *et al.*, 1981; Anand *et al.*, 1988; Hardy & Poteete, 1991). These two residues are solvent-accessible, but not generally replaceable. The interior location of residues that are sensitive to substitution is illustrated in Figure 4.

Perutz *et al.* (1965) observed that surface residues exhibited little conservation among naturally occurring globins with highly conserved three-dimensional structures. In general, they could be substituted by both polar and non-polar residues; polar residues, though, were excluded from interior positions. Sauer and co-workers, using combinatorial cassette mutagenesis, have found that positions on the surface, but not in the interior, of the amino-terminal domain of λ repressor, could tolerate substitution with hydrophilic residues (for a

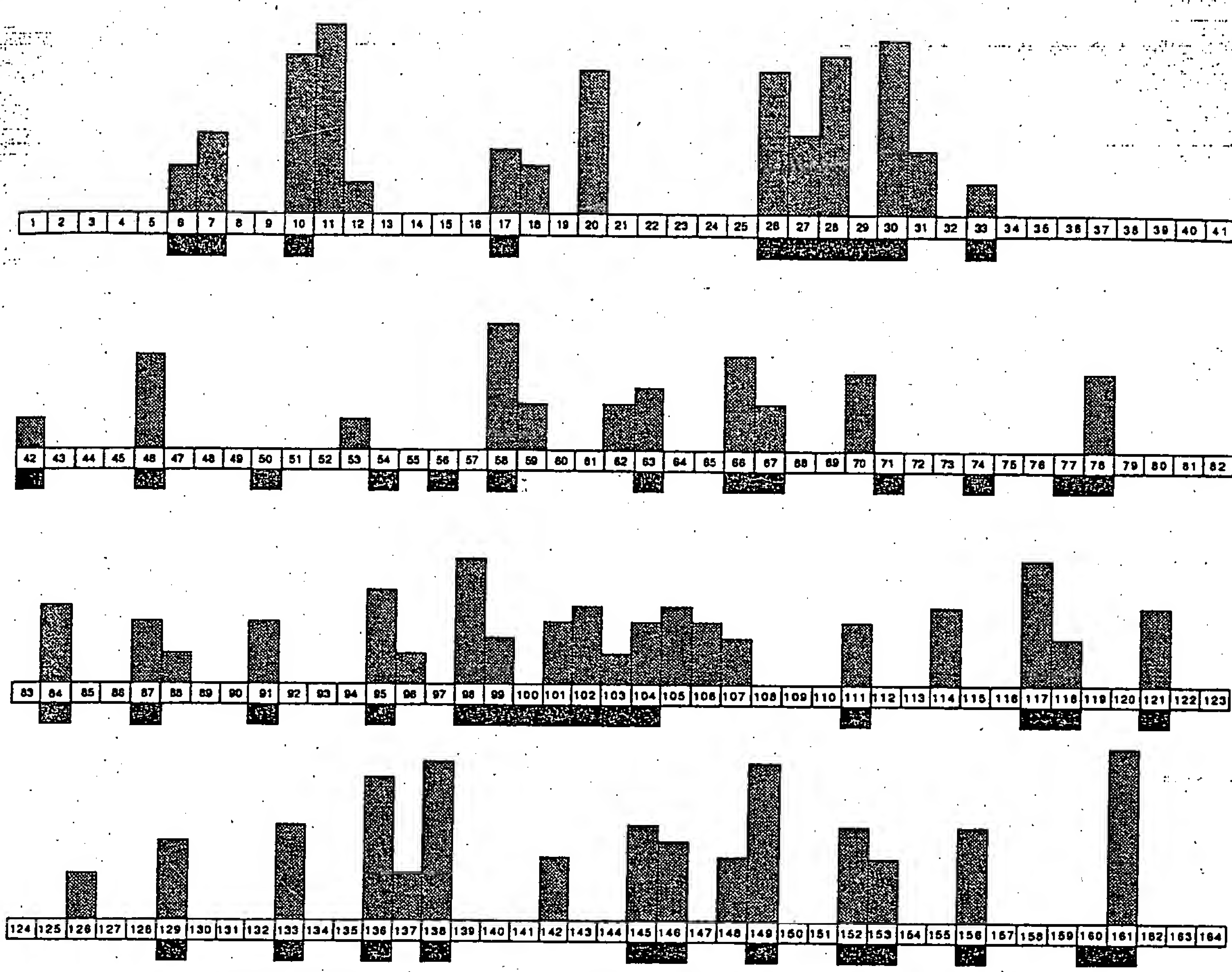


Figure 3. Sensitivity to substitution *versus* solvent inaccessibility. Positions of residues in the lysozyme sequence are indicated along the line by numbers. For each position, the height of the bar above the line is proportional to the number of deleterious substitutions found (Table 1), ranging from 2 to 12 (positions at which only 1 deleterious substitution occurs are not shown). Those residues with side-chain solvent accessibilities of less than 12% are indicated by bars below the line. (Solvent accessibility here refers to the calculated surface area that can be contacted by a sphere of radius 1·5 Å, expressed as a percentage of the accessible surface area of the same residue in the unfolded state: S. Dao-Pin & L. Weaver, personal communication.)

review, see Bowie *et al.*, 1990). In Table 2, each amino acid nominally inserted by a suppressor is scored according to the efficiency with which it replaces solvent-accessible and buried residues. As can be seen, all residues function nearly equally well in solvent-accessible positions, while Glu, Pro, Lys and Arg stand out as not being tolerated in buried positions. The conclusion that emerges from this analysis is that charged residues and proline are generally not acceptable in interior positions in lysozyme. The possibly surprising observation is that non-charged residues of high or moderate hydrophobicity, Gln in particular, generally are acceptable; this contrasts with what has been found for λ repressor. One possible explanation is that a substitution that buries a non-charged but polar residue in the hydrophobic core may, indeed, destabilize lysozyme, but not enough to be scored as deleterious in this system. That mutations must be highly deleterious to be scored as such in this system is clear from comparison of the data in Table 1 with accounts of detailed studies of mutant T4 lysozymes by others (see below).

(d) *Temperature sensitivity*

Of the 2015 substitutions indicated in Table 1, 261 exhibit qualitatively better function at 25°C than at 37°C. In 96 cases, this difference results in a crossing of the boundary between + and ±, or a classic temperature-sensitive phenotype. Temperature sensitivity of a mutant protein relative to wild-type is generally regarded as an indication of structural destabilization. Alber *et al.* (1987), in a study of 25 mutant T4 lysozymes, found that temperature-sensitive mutations occurred at sites with low mobility and low solvent accessibility. A striking correlation of the sites of the mutations with low values of average side-chain thermal factors in the refined structure model was noted. Relating the set of mutant proteins studied by these investigators to the entire set in Table 1 is complicated: the latter contains many of the former but, in addition, contains others with far more drastic effects on the protein, probably including many mutations that so destabilize lysozyme that it is not functional at any physiological temperature.

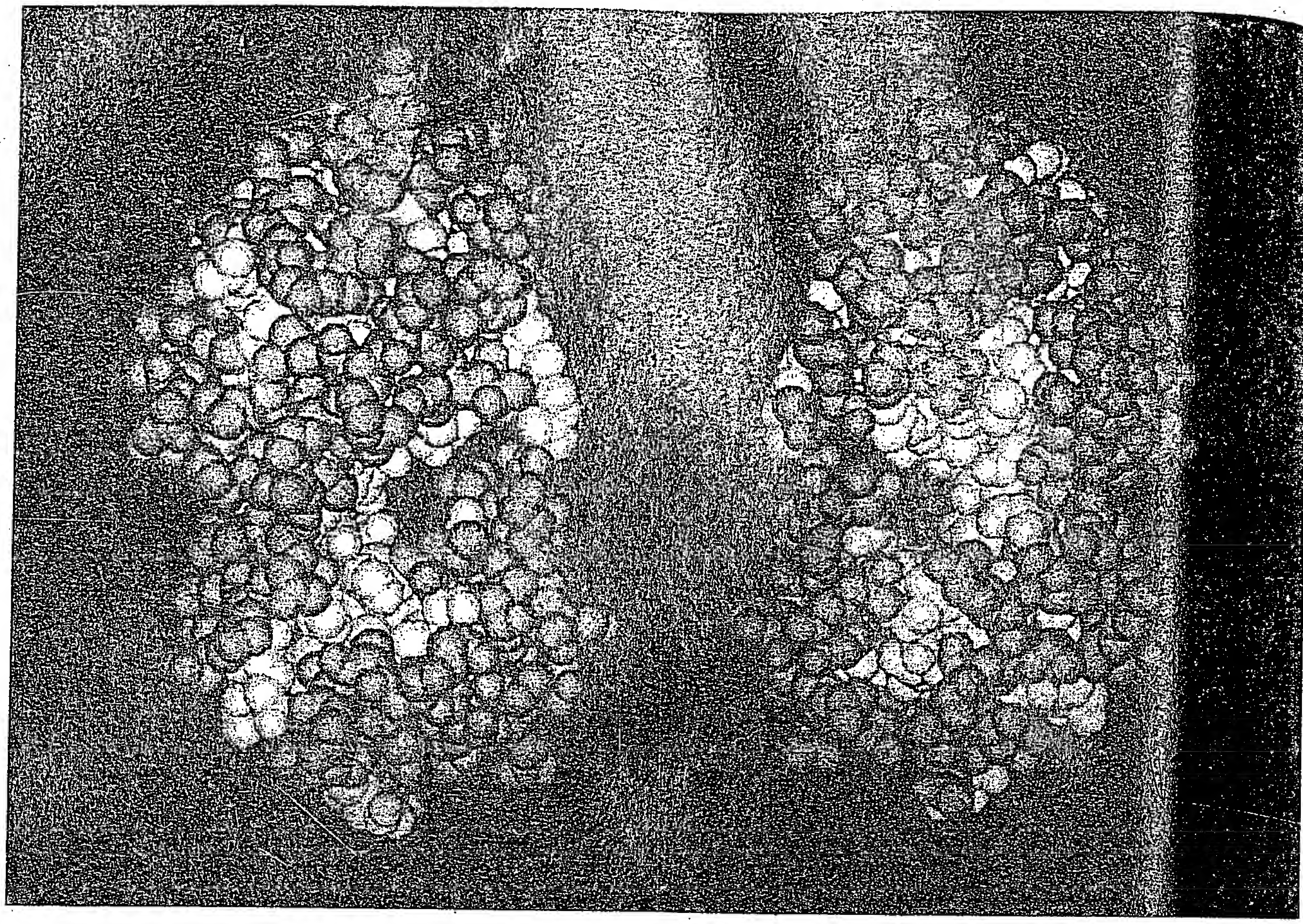


Figure 4. Positions of critical residues in T4 lysozyme. Atoms in white are those of residues whose replacement (by at least 2 others) leads to loss of function. The Figure was generated with Promodeler molecular graphics software (New England Biographics), using co-ordinates from the Brookhaven Protein Data Bank. Two views, rotated by 180°, are shown.

However, one prediction related to the findings of Alber *et al.* is that temperature-sensitive mutations would not be found at positions with high mobility. This is, indeed, the case, as shown in Figure 5, in which the tendency of substitutions at each position to exhibit temperature sensitivity is plotted as a function of average side-chain thermal factors. A statistical analysis of these data rejects the null hypothesis: that side-chain mobility is unrelated to the potential temperature sensitivity of its substitutions. Among 163 residues, temperature-sensitive substitutions are found in 42, or ~26%. The 37 residues with the highest thermal factors are not represented among these 42. The probability of picking, at random, a collection of 37 residues with no temperature-sensitive substitutions, would be ~0·000016.

The data in Table 1 also indicate 202 substitutions with better scores at 37°C than at 25°C, including 60 that cross the boundary between + and ±. It is uncertain whether any of these represent genuine cold-sensitive proteins. More than half of them occur on the Cys-inserting amber suppressor strain. The Cys suppressor strain grows well at 25°C, but does not suppress amber mutations well at this temperature, including ones at the two Cys codons in the lysozyme gene. In addition, the Tyr suppressor strain is itself mildly cold-sensitive; it grows poorly at 25°C. Moreover, all the apparently cold-sensitive phenotypes are cases in which the scores are + at 37°C and ± at 25°C; no greater differences are seen. Finally, the hybrid phages lack P22 gene *15*. Although this gene is not essential, its small contribution to P22's plaque-forming ability is relatively greater at low temperatures (Casjens *et al.*, 1989); thus, we might expect that the effect of a mild mutation in the lysozyme gene of P22 *e416* would be slightly exacerbated at 25°C relative to 37°C, unless the mutant protein itself were stabilized by the lower temperature.

(e) *Proline-sensitive α-helix*

The inability of proline residues to adopt many of the backbone conformations available to amino acid residues might impose severe limitations on the number of positions in a protein that it can occupy. In particular, proline residues cannot fit into an α-helix without distorting it, as well as destabilizing it due to its lack of a backbone amide for hydrogen

Table 2
Average suppression scores

	Positions in lysozyme	
	Solvent-exposed	Buried
Suppressors		
Hydrophobic†		
Ala	2·9	2·8
Cys	2·5	1·8
Leu	2·9	2·3
Phe	2·7	2·0
Moderately polar		
Ser	2·9	2·5
His	2·9	1·8
Tyr	2·8	1·9
Very polar		
Glu	2·7	1·0
Gln	2·9	2·2
Lys	2·4	1·0
Arg	2·7	0·9
Exceptional backbone torsional properties		
Gly	2·9	2·4
Pro	2·4	1·1

Data in Table 1 are converted to numbers (+ + = 3, + = 2, ± = 1, − = 0), and averages are calculated for different subsets of residues, defined as for Fig. 2.

† Residues are assigned to groups according to Rose *et al.* (1985), except that Gly and Pro are placed in a separate group.

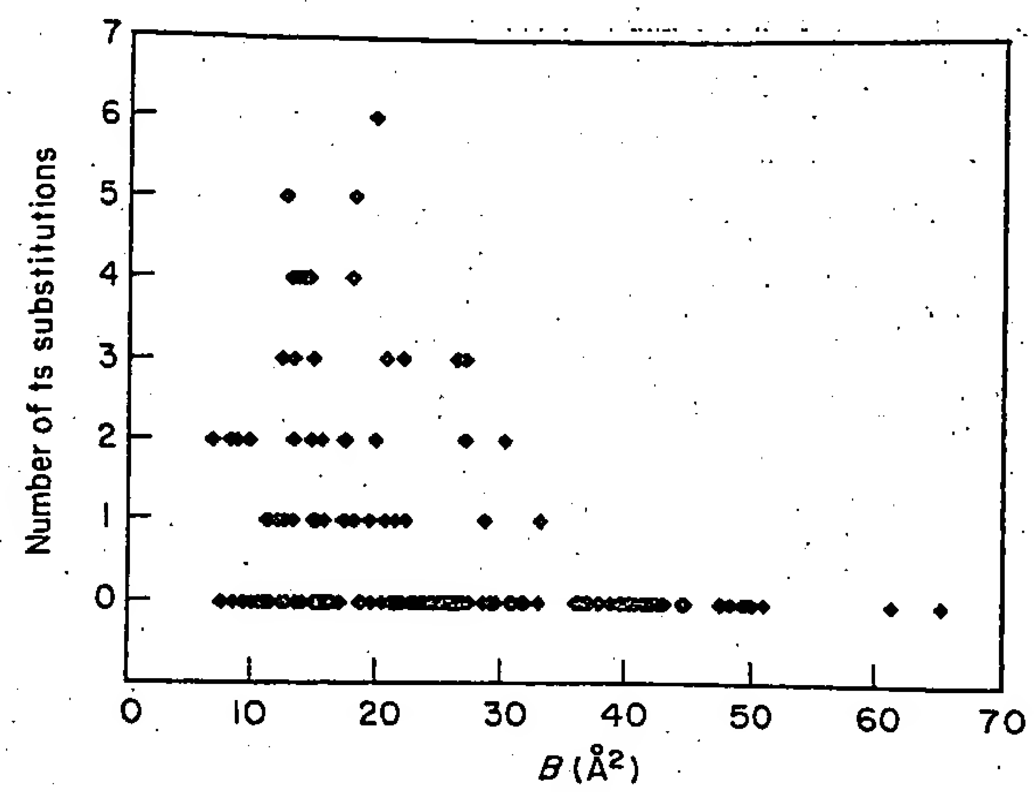


Figure 5. Temperature sensitivity (ts) *versus* side-chain mobility in the folded structure. The number of temperature-sensitive substitutions (those resulting in scores of + + or + at 25 °C, ± or − at 37 °C in Table 1) that occur at each residue is plotted as a function of the average thermal factor (B) of its side-chain atoms (not including α-carbon atoms, except in the case of Gly residues).

bonding. Indeed, the data in Table 1 show that proline is the second most frequently unacceptable residue; 53 proline substitutions are deleterious, compared with 58 lysine substitutions, the most frequently unacceptable. If only the strongest defects, those resulting in a (−) score, are considered, proline is the most frequently unacceptable residue. In contrast, only four alanine substitutions are deleterious. The pattern of proline-sensitive positions in the polypeptide does not particularly reflect the locations of α-helices, however. T4 lysozyme contains nine α-helices, comprising residues 3 to 10, 39 to 49, 60 to 79, 82 to 90, 95 to 106, 115 to 122, 126 to 134, 137 to 141 and 143 to 155 (Remington *et al.*, 1978). Of these, only one, 95 to 106, shows a marked sensitivity to proline substitutions (except in the first turn, where proline residues are often found in α-helices). This particular α-helix is also unique in that a large part of it is completely internal, running through the core of the carboxy-terminal domain; the others, for the most part, lie on the surface of the molecule. Proline substitutions in this α-helix may have more profoundly destabilizing effects than usual because of its interior location. This hypothesis is supported by the dependence upon solvent inaccessibility of any α-helical residue's sensitivity to proline substitution. Of the 26 α-helical residues which are >95% buried, 19 are sensitive. This contrasts sharply with the observation that only two of the 37 α-helical residues which are <50% buried are sensitive. This greater relative sensitivity of buried α-helical residues suggests that distortion of the helical conformation by proline substitution is, alone, insufficient to inactivate the protein; in most cases, further destabilization by the loss of a hydrogen bond donor for a buried backbone carbonyl is required. This idea is supported by consideration of those seven exceptional α-helical residues which are >95% buried but not sensitive to proline substitution. In each case the peptide nitrogen is bonded to a carbonyl oxygen that is capable of forming a hydrogen bond with an alternative donor, usually solvent water.

(f) *Critical residues*

The data in Table 1 suggest a rough rank order of residues by sensitivity to substitution. The 12 positions with the lowest total suppression scores (counting + + as 3, + as 2, ± as 1, and − as 0) are Glu11 (score: 3), Gly30 (7), Tyr161 (7), Asp10 (8), Trp138 (8), Val149 (9), Gly28 (10), Ser136 (14), Thr26 (15), Ala98 (15), Asp20 (16) and Ile58 (18). The crystal structure of T4 lysozyme (Weaver & Matthews, 1987), as well as other observations, in many cases suggest plausible explanations for these sensitivities.

Glu11 and Asp20 are assumed to be directly involved in catalysis on the basis of structural homology with their counterparts in hen egg white lysozyme, Glu35 and Asp52 (Anderson *et al.*, 1981). Anand *et al.* (1988) have shown that substitutions of Gln and Asn, respectively, at these two positions lead to loss of enzymatic activity. The extreme sensitivity of Glu11 to substitution is fully consistent with its designation as the key catalytic residue. Asp20 can be replaced by Glu, as found by Anand *et al.* (1988); surprisingly, though, it can evidently be replaced by either Cys or Ala as well. We have purified T4 lysozyme bearing an Asp20→Cys substitution (as the result of the direct

substitution of a Cys codon), and shown that it retains 80% of the specific activity of wild-type, and has, in addition, acquired a new sensitivity to thiol-modifying reagents (Hardy & Poteete, 1991). Similarly, P22 *e416* bearing an Ala sense codon at position 20 forms plaques well, indicating that the result with the amber suppressor is not misleading. Moreover, if the sequence of the ϕ29 lysozyme is aligned by homology with T4 lysozyme, it is found to have an alanine residue at the position occupied by Asp20 in T4 lysozyme (Garvey *et al.*, 1986). These results would seem to call into question the nature or even existence of a catalytic role for Asp20 (Hardy & Poteete, 1991), but further studies are needed to clarify this issue.

T4 lysozyme contains two buried salt bridges, between Arg145 and Glu11, as well as Arg148 and Asp10. All of these residues could conceivably be important for stabilizing the position of the key catalytic residue Glu11. Indeed, Asp10 and Arg145 are among the residues most sensitive to substitution, and Arg148 is sensitive to at least some substitutions. On the other hand, the data in Table 1 do not indicate a stringent requirement for salt bridges *per se*, because all but Glu11 can be replaced successfully with uncharged residues (but note that only in the case of Arg148 do these substitutions result in no diminution of plaque size).

The two buried salt bridges are themselves part of a larger network of internal interactions among a cluster of amino acid side-chains in the carboxy-terminal domain. One of the key residues in this cluster is Tyr161. Its hydroxyl makes a hydrogen bond to the carboxyl group of Asp10, and closely approaches the side-chain of Val149; its aromatic ring fits neatly into a pocket created by the side-chains of Met1 and Met6. The amide nitrogen of Asn101 makes an additional hydrogen bond with the carboxyl group of Asp10, and Asn101, in turn, also closely approaches the side-chain of Val149. The functional importance of this network is clearly shown by the data in Table 1, which indicate that all of these residues (except for Met1, which was not tested), are not freely replaceable; indeed, positions 161 and 149 are among the most sensitive to substitution in the molecule.

Among the other positions most sensitive to substitution, three, Gly30, Gly28 and Thr26, are small residues lining the active site cleft. All three can be most effectively replaced with other small residues. It seems plausible that the presence of bulky side-chains at these positions would interfere with substrate binding. Consistent with this idea, a mutant T4 lysozyme bearing a substitution of Gln for Thr26 is nearly unaltered in thermal stability, but highly defective in enzymatic activity (Poteete *et al.*, 1991).

The hydroxyl group of Ser136 is buried, and hydrogen-bonded to the backbone amide nitrogen of Tyr139. Substitution by larger residues at this position is generally not tolerated. In the resulting mutant proteins, the polypeptide backbone at this position would most likely bulge out, perhaps making the molecule unstable or sensitive to proteases *in vivo*.

The methyl side-chain of Ala98 fits into a small space bounded by a number of backbone atoms, particularly closely by those of residues 99, 95 and 149. Substitution of Ala98 by any amino acid residue with a larger side-chain might be expected to lead to a significant structural alteration because of steric clash. Such structural alterations have, in fact, been observed in lysozyme bearing an Ala98$\rightarrow$Val substitution (Alber & Matthews, 1987). Consistent with this concept, the data in Table 1 indicate that substitutions by residues larger than Cys are not tolerated, with the exception of Pro, which is.

The ring nitrogen of Trp138 is hydrogen-bonded to the side-chain amide oxygen of Gln105. Otherwise, there is little to distinguish Trp138 from any of the other residues of the hydrophobic core of the molecule, and thus to account for its extraordinary sensitivity to substitutions, which greatly exceeds that of the other hydrophobic core residues. Only Phe, Tyr and Leu are tolerated at this position.

The data in Table 1 indicate that Ile58 is sensitive to substitution. This residue is part of the hydrophobic core of the N-terminal domain. The pattern of its sensitivity to replacement by other residues is generally consistent with that of other hydrophobic core residues, but is exceptionally stringent: the phage bearing an amber mutation in codon 58 does not form plaques on the leucine-inserting amber suppressor strain. Examination of the structure does not immediately suggest why this position should be sensitive to such a conservative substitution.

(g) *Exceptional residues*

Figure 3 indicates a number of residues that are solvent-exposed, but sensitive to several different substitutions. In addition to the catalytic Glu11 and Asp20, these include His31, Asp70, Gln105, Phe114 and Thr142. The sensitivities to substitution of Gln105 and Thr142 are readily understandable in light of the model of Anderson *et al.* (1981): both are likely involved in contacts with substrate.

His31 and Asp70 are salt-bridged to each other. Anderson *et al.* (1990) have studied mutant T4 lysozymes in which the His31–Asp70 salt bridge is disrupted by replacing these residues with Asn, and concluded that it contributes significantly to the thermal stability of the protein. Although the data in Table 1 do not include Asn substitutions, they indicate that both positions are sensitive to a number of other substitutions. It would seem reasonable to attribute the defects of the substituted proteins to destabilization *via* disruption of the salt bridge, except that some of the fully functional substitutions at these positions involve residues that cannot participate in salt bridges, or even hydrogen bonds. Possibly, these residues

contribute to lysozyme function in other ways as well.

Phe114 sits in a crevice on the surface of the C-terminal domain, with one face of the aromatic ring packed against the hydrophobic interior, and the other fully exposed to solvent, yet it is as sensitive to substitutions as residues that constitute the hydrophobic core. Phe114 has, in addition, not been implicated in substrate binding or catalysis (Anderson *et al.*, 1981). One hypothesis to account for its apparently critical role in the protein is that it normally engages in an aromatic–aromatic interaction, of the type described by Burley & Petsko (1985), with Trp138. Understanding of the role of Phe114 will have to await characterization of the properties of mutant proteins altered at this position.

Results summarized in Figure 3 and Table 1 indicate that five relatively solvent-inaccessible residues are completely insensitive to substitutions: Cys54, Gly56, Val71, Ala74 and Gly77. The side-chain of Cys54 projects into the interior of the amino-terminal domain. Replacing it with large residues would appear to require some local rearrangement of the structure; its insensitivity to substitutions is thus not readily apparent from the structure. On the other hand, accommodating a large side-chain in the position of Gly56 would appear to require only displacing the side-chain of either Lys16 or Lys43, both of which lie along the surface of the protein. In the case of Val71, one of the γ-carbon atoms is exposed to solvent; it may be possible to extend the longer side-chains of large residues distally from this position without greatly distorting the structure. Similarly, the β-carbon of Ala74 is solvent-accessible from the direction in which a larger side-chain would extend. Access of the α-carbon of Gly77 to solvent is blocked only by the side-chain of Arg80, a residue with high side-chain mobility lying along the surface of the lysozyme molecule.

(h) *Relationship of these results to previous studies*

T4 lysozyme has been the subject of genetic studies for many years (see Streisinger *et al.*, 1966); many mutant lysozymes have been characterized. The present systematic study of single amino acid substitutions effected by amber suppression thus, directly or indirectly, reproduces parts of many previous studies.

The results shown in Table 1 directly reproduce those of Tsugita and co-workers (for a review, see Tsugita, 1971), who determined the suppression patterns of amber mutants affecting codons 126, 138 and 158 (all normally encoding Trp), as well as codons 69, 105, 122, 123 and 141 (all Gln residues) on amber suppressors inserting Ser, Gln and Tyr. Although these investigators examined the plating of phage T4 on *E. coli*, while we tested the plating of a hybrid P22 phage on *S. typhimurium*, our observations are in agreement with theirs.

Most of the characterized mutant T4 lysozymes result from missense mutations, or from combinations of frameshift mutations that restore the reading frame but cause deletions, insertions and substitutions. Characterization of the latter class was historically important in establishing the nature of the genetic code (see Streisinger *et al.*, 1966).

Remington *et al.* (1978) summarized the phenotypes of most of the then existing, characterized T4 lysozyme mutants and interpreted them in light of the crystal structure. Phenotypically mild mutants of T4 lysozyme bearing multiple substitutions as a result of compensating frame-shift mutations occur in stretches of residues that are relatively insensitive to substitutions, according to the data in Table 1. These stretches include 2 to 4, 22 to 25, 34 to 40, 73 to 76 and 139 to 140. On the other hand, more seriously defective mutants of the same derivation affect residues identified in Table 1 as sensitive to substitution: Asp20, Trp126 and Trp138. In further agreement with the results in Table 1, these authors reported that the substitutions Asn2→Arg, Thr34→Gln, Tyr88→His and Gln141→Arg all have relatively small effects on T4 lysozyme activity. Moreover, they reported that removal of the last two residues of T4 lysozyme with carboxypeptidase had little effect on the enzyme; our finding that amber mutations in the last three codons are non-deleterious in the non-suppressing host is consistent with this result.

The lysozyme of the related phage T2 differs in sequence from that of T4 by three single amino acid substitutions: Asn40→Ser, Ala41→Val and Thr151→Ala (Inouye & Tsugita, 1968). The two of these changes that are represented in Table 1 have no effect on lysozyme activity, as would be expected. The third, Ala41→Val, has been shown to increase the thermal stability of T4 lysozyme (Dao-Pin *et al.*, 1990).

Perry & Wetzel (1985) described a mutant T4 lysozyme, Ile3→Cys, in which a novel disulfide bond was formed between Cys3 and Cys97. The results shown in Table 1 suggest that this mutant protein is active, in agreement with the findings of these investigators. These investigators also found (Perry & Wetzel, 1987) that substitutions of Val and Ser for Cys54 and Cys97, respectively, stabilize the protein to oxidative damage, and do not harm activity. The latter mutation is represented in Table 1, and is scored as functional. Similarly, Matthews and co-workers have used a double mutant, Cys54→Thr/Cys97→Ala, as a starting point for studies of mutant variants, because this "cysteine-less wild-type" is not as sensitive to oxidation as the wild-type (Matsumura & Matthews, 1989; Pjura *et al.*, 1990).

Alber & Matthews (1987) described a set of 21 mutant lysozymes with single amino acid substitutions as falling into four classes: tight ts, leaky ts, low activity and heat resistant-low activity. Of these 21 mutants, 12 are represented in Table 1. Both members of the tight ts class, Leu66→Pro and Leu91→Pro, are defective in our system

(ts = temperature-sensitive); the latter exhibits a particularly pronounced temperature sensitivity. Among the substitutions classified as leaky ts, eight are represented in Table 1; only one, Trp126→Arg, is scored as defective. The single low activity mutant Glu128→Lys (originally described by Grutter & Matthews, 1982) is not scored as deleterious in Table 1, nor is the single heat-resistant low activity mutant Cys54→Tyr. Alber & Matthews (1987) additionally described a set of 13 single amino acid substitutions of Thr157, all of which mildly destabilize the protein. Table 1 nominally contains information on nine of these substitutions; none is defective in function. Alber *et al.* (1988) further described a series of ten single amino acid substitutions of Pro86, all of which had small effects on stability and activity. Eight of these substitutions are represented in Table 1; none is defective. Among the 25 temperature-sensitive substitutions described by Alber *et al.* (1987), 14 are represented in Table 1; five of them (including one not previously mentioned, Trp138→Gly) are scored as defective.

Matsumura *et al.* (1988) described a series of 13 single amino acid substitutions of Ile3, all of which had mild effects on protein stability. None of the eight of these substitutions represented in Table 1 is scored as defective there. Other T4 lysozyme mutants characterized by Matthews and co-workers as having little or no effect on activity or stability, and which are additionally represented in Table 1, include: Leu133→Phe (Karpusas *et al.*, 1989); Gly77→Ala and Ala82→Pro (Matthews *et al.*, 1987); Asn55→Gly and Lys124→Gly (Nicholson *et al.*, 1989); and Val131→Ala (Dao-Pin *et al.*, 1990). All of these are scored as fully functional in Table 1.

Overall, the results shown in Table 1 are in good agreement with previous studies of T4 lysozyme mutants, including a number of unpublished ones (S. Dao-Pin, E. Eriksson, A. Morton & B. Matthews, personal communications). The best-characterized mutant lysozymes generally have minimally altered structures. In general, such mutant proteins are not scored as defective in Table 1, even though some of them have significantly reduced thermal stabilities. However, it is reasonable to assume that a mutation that reduces the melting temperature of T4 lysozyme at nearly neutral pH from over 60°C to around 50°C (for instance) would not inactivate lysozyme function at 37°C. Thus, in general, the functional test employed in these studies is stringent: only strong mutations are scored as deleterious.

We thank Brian Matthews, Larry Weaver, Andrew Morton, Cai Zhang, Sun Dao-Pin, Elisabeth Eriksson and Eric Anderson for communicating unpublished results on T4 lysozyme; and Robert Sauer for comments on the manuscript. We thank J. Miller and W. McClain for supplying amber suppressor-bearing plasmids. We thank Jeff Barbon for technical assistance. This research was supported by grant AI24083 from NIH.

References

Alber, T. & Matthews, B. W. (1987). Structure and thermal stability of phage T4 lysozyme. *Methods Enzymol.* **154**, 511–533.

Alber, T., Dao-Pin, S., Nye, J. A., Muchmore, D. C. & Matthews, B. W. (1987). Temperature-sensitive mutations of bacteriophage T4 lysozyme occur at sites with low mobility and low solvent accessibility in the folded protein. *Biochemistry*, **26**, 3754–3758.

Alber, T., Bell, J. A., Dao-Pin, S., Nicholson, H., Wozniak, J. A., Cook, S. & Matthews, B. W. (1988). Replacements of Pro86 in phage T4 lysozyme extend an alpha-helix but do not alter protein stability. *Science*, **239**, 631–635.

Anand, N. N., Stephen, E. R. & Narang, S. A. (1988). Mutation of active site residues in synthetic T4-lysozyme gene and their effect on lytic activity. *Biochem. Biophys. Res. Commun.* **153**, 862–868.

Anderson, D. E., Becktel, W. J. & Dahlquist, F. W. (1990). pH-induced denaturation of proteins: a single salt bridge contributes 3–5 kcal/mol to the free energy of folding of T4 lysozyme. *Biochemistry*, **29**, 2403–2408.

Anderson, W. F., Grutter, M. G., Remington, S. J., Weaver, L. H. & Matthews, B. W. (1981). Crystallographic determination of the mode of binding of oligosaccharides to T4 bacteriophage lysozyme: implications for the mechanism of catalysis. *J. Mol. Biol.* **147**, 523–543.

Bebenck, K. & Kunkel, T. A. (1989). The use of native T7 DNA polymerase for site-directed mutagenesis. *Nucl. Acids Res.* **17**, 5408.

Bossi, L. (1983). Context effects: translation of UAG codon by suppressor tRNA is affected by the sequence following UAG in the message. *J. Mol. Biol.* **164**, 73–87.

Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A. & Sauer, R. T. (1990). Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science*, **247**, 1306–1310.

Brent, R. & Ptashne, M. (1981). Mechanism of action of the *lexA* gene product. *Proc. Nat. Acad. Sci., U.S.A.* **78**, 4204–4208.

Burley, S. K. & Petsko, G. A. (1985). Aromatic–aromatic interaction: a mechanism of protein structure stabilization. *Science*, **229**, 23–28.

Casjens, S., Eppler, K., Parr, R. & Poteete, A. R. (1989). Nucleotide sequence of the bacteriophage P22 gene *19* to *3* region: identification of a new gene required for lysis. *Virology*, **171**, 588–598.

Dao-Pin, S., Baase, W. A. & Matthews, B. W. (1990). A mutant T4 lysozyme (Val131→Ala) designed to increase thermostability by the reduction of strain within an α-helix. *Proteins*, **7**, 198–204.

Garvey, K. J., Saedi, M. S. & Ito, J. (1986). Nucleotide sequence of *Bacillus* phage ϕ29 genes *14* and *15*: homology of gene *15* with other phage lysozymes. *Nucl. Acids Res.* **14**, 10001–10008.

Grana, D., Gardella, T. & Susskind, M. M. (1988). The effects of mutations in the *ant* promoter of phage P22 depend on context. *Genetics*, **120**, 319–327.

Grutter, M. G. & Matthews, B. W. (1982). Amino acid substitutions far from the active site of bacteriophage T4 lysozyme reduce catalytic activity and suggest that the C-terminal lobe of the enzyme participates in substrate binding. *J. Mol. Biol.* **154**, 525–535.

Hardy, L. W. & Poteete, A. R. (1991). Re-examination of

the role of Asp20 in catalysis by bacteriophage T4 lysozyme. *Biochemistry*, in the press.

Inoue, T. & Cech, T. R. (1985). Secondary structure of the circular form of the *Tetrahymena* rRNA intervening sequence: a technique for RNA structure analysis using chemical probes and reverse transcriptase. *Proc. Nat. Acad. Sci., U.S.A.* **82**, 648–652.

Inouye, M. & Tsugita, A. (1968). Amino acid sequence of T2 phage lysozyme. *J. Mol. Biol.* **37**, 213–223.

Karpusas, M., Baase, W. A., Matsumura, M. & Matthews, B. W. (1989). Hydrophobic packing in T4 lysozyme probed by cavity-filling mutants. *Proc. Nat. Acad. Sci., U.S.A.* **86**, 8237–8241.

Kleina, L. G. & Miller, J. H. (1990). Genetic studies of the *lac* repressor XIII. Extensive amino acid replacements generated by the use of natural and synthetic nonsense suppressors. *J. Mol. Biol.* **212**, 295–318.

Kleina, L. G., Masson, J.-M., Normanly, J., Abelson, J. & Miller, J. H. (1990). Construction of *Escherichia coli* amber suppressor tRNA genes. II. Synthesis of additional tRNA genes and improvement of suppressor efficiency. *J. Mol. Biol.* **213**, 705–717.

Knight, J. A., Hardy, L. W., Rennell, D., Herrick, D. & Poteete, A. R. (1987). Mutations in an upstream regulatory sequence that increase expression of the bacteriophage T4 lysozyme gene. *J. Bacteriol.* **169**, 4630–4636.

Kramer, W. & Fritz, H.-J. (1987). Oligonucleotide-directed construction of mutations *via* gapped duplex DNA. *Methods Enzymol.* **154**, 350–367.

Loeb, D. D., Swanstrom, R., Everitt, L., Manchester, M., Stamper, S. E. & Hutchison, C. A., III (1989). Complete mutagenesis of the HIV-1 protease. *Nature (London)*, **340**, 397–400.

Matsumura, M. & Matthews, B. W. (1989). Control of enzyme activity by an engineered disulfide bond. *Science*, **243**, 792–794.

Matsumura, M., Becktel, W. J. & Matthews, B. W. (1988). Hydrophobic stabilization in T4 lysozyme determined directly by multiple substitutions of Ile3. *Nature (London)*, **334**, 406–410.

Matthews, B. W. (1987). Genetic and structural analysis of the protein stability problem. *Biochemistry*, **26**, 6885–6888.

Matthews, B. W., Nicholson, H. & Becktel, W. J. (1987). Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding. *Proc. Nat. Acad. Sci., U.S.A.* **84**, 6663–6667.

McClain, W. H. & Foss, K. (1988). Changing the acceptor identity of a transfer RNA by altering nucleotides in a "variable pocket". *Science*, **241**, 1804–1807.

Miller, J. H. & Albertini, A. M. (1983). Effects of surrounding sequence on the suppression of nonsense codons. *J. Mol. Biol.* **164**, 59–71.

Miller, J. H., Coulondre, C., Hofer, M., Schmeissner, U., Sommer, H., Schmitz, A. & Lu, P. (1979). Genetic studies of the *lac* repressor IX. Generation of altered proteins by the suppression of nonsense mutations. *J. Mol. Biol.* **131**, 191–222.

Nicholson, H., Soderlind, E., Tronrud, D. E. & Matthews, B. W. (1989). Contributions of left-handed helical residues to the structure and stability of bacteriophage T4 lysozyme. *J. Mol. Biol.* **210**, 181–193.

Normanly, J., Masson, J.-M., Kleina, L. G., Abelson, J. & Miller, J. H. (1986). Construction of two *Escherichia coli* amber suppressor genes: $tRNA^{Phe}_{CUA}$ and $tRNA^{Cys}_{CUA}$. *Proc. Nat. Acad. Sci., U.S.A.* **83**, 6548–6552.

Normanly, J., Kleina, L., Masson, J.-M., Abelson, J. & Miller, J. H. (1990). Construction of *Escherichia coli* amber suppressor tRNA genes. III. Determination of tRNA specificity. *J. Mol. Biol.* **213**, 719–726.

Perry, L. J. & Wetzel, R. (1984). Disulfide bond engineered into T4 lysozyme; stabilization of the protein toward thermal inactivation. *Science*, **226**, 555–557.

Perry, L. J. & Wetzel, R. (1987). The role of cysteine oxidation in the thermal inactivation of T4 lysozyme. *Protein Eng.* **1**, 101–105.

Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965). Structure and function of haemoglobin. II. Some relations between polypeptide chain configuration and amino acid sequence. *J. Mol. Biol.* **13**, 669–678.

Phipps, J., Michniewicz, J., Yao, F.-L. & Narang, S. A. (1987). Retention of enzymatic activity by N-terminal domain (1–78) T4-lysozyme: expression of synthetic DNA in *Escherichia coli*. *Biochem. Biophys. Res. Commun.* **145**, 190–195.

Pjura, P. E., Matsumura, M., Wozniak, J. A. & Matthews, B. W. (1990). Structure of a thermostable disulfide-bridge mutant of phage T4 lysozyme shows that an engineered crosslink in a flexible region does not increase the rigidity of the folded protein. *Biochemistry*, **29**, 2592–2598.

Poteete, A. R., Dao-Pin, S., Nicholson, H. & Matthews, B. W. (1991). Second-site revertants of an inactive T4 lysozyme mutant restore activity by restructuring the active site cleft. *Biochemistry*, **30**, 1425–1432.

Reidhaar-Olson, J. F. & Sauer, R. T. (1988). Combinatorial cassette mutagenesis as a probe of the informational content of protein sequences. *Science*, **241**, 53–57.

Remington, S. J., Anderson, W. F., Owen, J., Ten-Eyck, L. F., Grainger, C. T. & Matthews, B. W. (1978). Structure of the lysozyme from bacteriophage T4: an electron density map at 2·4 Å resolution. *J. Mol. Biol.* **118**, 81–98.

Rennell, D. & Poteete, A. R. (1989). Genetic analysis of bacteriophage P22 lysozyme structure. *Genetics*, **123**, 431–440.

Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H. & Zehfus, M. H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science*, **229**, 834–838.

Russel, M., Kidd, S. & Kelley, M. R. (1986). An improved filamentous helper phage for generating single-stranded plasmid DNA. *Gene*, **45**, 333–338.

Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzaghi, E. & Inouye, M. (1966). Frameshift mutations and the genetic code. *Cold Spring Harbor Symp. Quant. Biol.* **31**, 77–84.

Tsugita, A. (1971). Phage lysozyme and other lytic enzymes. In *The Enzymes* (Boyer, P. D., ed.), vol. 5, 3rd edit., pp. 343–411, Academic Press, New York.

Tsugita, A. & Inouye, M. (1968). Complete primary structure of phage lysozyme from *Escherichia coli* T4. *J. Mol. Biol.* **37**, 201–212.

Weatherall, D. J. & Clegg, J. B. (1976). Molecular genetics of human hemoglobin. *Annu. Rev. Genet.* **10**, 157–178.

Weaver, L. H. & Matthews, B. W. (1987). Structure of bacteriophage T4 lysozyme refined at 1·7 Å resolution. *J. Mol. Biol.* **193**, 189–199.

Winston, F., Botstein, D. & Miller, J. H. (1979).

Characterization of amber and ochre suppressors in *Salmonella typhimurium*. *J. Bacteriol.* **137**, 433–439.
Youderian, P., Bouvier, S. & Susskind, M. M. (1982). Sequence determinants of promoter activity. *Cell*, **30**, 843–853.
Zagursky, R. J. & Berman, M. L. (1984). Cloning vectors that yield high levels of single-stranded DNA for rapid DNA sequencing. *Gene*, **27**, 183–191.

Edited by P. von Hippel

Protein tolerance to random amino acid change

Haiwei H. Guo*, Juno Choe†, and Lawrence A. Loeb*‡

*Joseph Gottstein Memorial Cancer Laboratory, Departments of Pathology and Biochemistry, University of Washington School of Medicine, Seattle, WA 98195-7705; and †Institute for Systems Biology, Seattle, WA 98103

Communicated by Leroy E. Hood, Institute for Systems Biology, Seattle, WA, May 10, 2004 (received for review March 23, 2004)

Mutagenesis of protein-encoding sequences occurs ubiquitously; it enables evolution, accumulates during aging, and is associated with disease. Many biotechnological methods exploit random mutations to evolve novel proteins. To quantitate protein tolerance to random change, it is vital to understand the probability that a random amino acid replacement will lead to a protein's functional inactivation. We define this probability as the "*x* factor." Here, we develop a broadly applicable approach to calculate *x* factors and demonstrate this method using the human DNA repair enzyme 3-methyladenine DNA glycosylase (AAG). Three gene-wide mutagenesis libraries were created, each with 10^5 diversity and averaging 2.2, 4.6, and 6.2 random amino acid changes per mutant. After determining the percentage of functional mutants in each library using high-stringency selection (>19,000-fold), the *x* factor was found to be 34% ± 6%. Remarkably, reanalysis of data from studies of diverse proteins reveals similar inactivation probabilities. To delineate the nature of tolerated amino acid substitutions, we sequenced 244 surviving AAG mutants. The 920 tolerated substitutions were characterized by substitutability index and mapped onto the AAG primary, secondary, and known tertiary structures. Evolutionarily conserved residues show low substitutability indices. In AAG, β strands are on average less substitutable than α helices; and surface loops that are not involved in DNA binding are the most substitutable. Our results are relevant to such diverse topics as applied molecular evolution, the rate of introduction of deleterious alleles into genomes in evolutionary history, and organisms' tolerance of mutational burden.

A fundamental aspect of evolution is that mutations generate novel alleles that are then favored by selection. However, new coding mutations can be deleterious, neutral, or beneficial. Mutations can result from environmental and endogenous damage to DNA and from errors during DNA synthetic processes. In humans, random mutations produce inherited diseases and accumulate with aging and cancer (1). Conversely, targeted hypermutagenesis by immune defenses helps to generate antibody diversity and was recently shown to inactivate retroviral genomes (2). John Maynard Smith (3) proposed more than 30 years ago that the occurrence of functional mutant proteins that differ from wild type by one residue is likely frequent for evolution to be possible. Since then, numerous evolutionary and mutagenesis studies have led to the assertion that proteins are highly plastic in tolerating amino acid changes (4, 5). However, to date, we lack a quantitative measure of the degree of proteins' tolerance for random amino acid changes that occur at a random position in the protein. If a rigorous measure of proteins' degree of tolerance of random amino acid changes can be defined, then such fundamental calculations as the steepness of protein fitness landscapes or the rate of introduction of deleterious mutations into coding genomes can be more clearly delineated. Further understanding of the nature of tolerated amino acid substitutions can also lend insight into protein folding and design.

Here, we develop the concept of the probability of inactivating a protein with a random codon replacement producing amino acid change at a random location along its sequence. For conciseness, this concept is named the *x* factor. We describe an analytical method for calculating the *x* factor of proteins from randomly mutated libraries and demonstrate the method using the human DNA repair enzyme 3-methyladenine DNA glycosylase [AAG, methyl purine DNA glycosylase (MPG), and alkyl purine DNA glycosylase (ANPG)]. Nine hundred and twenty tolerated amino acid substitutions in active mutant enzymes were identified and substitutions were mapped to the available x-ray crystal structure of AAG. We examine the applicability of the *x* factor concept to diverse proteins by reanalyzing results from prior studies. These findings reveal a similar range of inactivation probabilities.

Materials and Methods

Escherichia coli strain MV1932 (*ada alkA1*) was previously derived from strain AB1157 (6). Chemicals were from Sigma-Aldrich, enzymes were from NEB (Beverly, MA), and DNA oligonucleotides were purchased from IDT (Coralville, IA), unless otherwise indicated.

Construction of PCR Mutagenesis Libraries. The low, medium, and highly mutated AAG libraries were generated by using a previously undescribed PCR mutagenesis protocol that produces similar mutational frequencies at G:C and A:T base pairs. Briefly, PCR mutagenesis was carried out sequentially with Mutazyme in the GENEMORPH kit (Stratagene), which preferentially mutates at G:C sites, and with *Taq*DNA polymerase with 0.5 mM Mn^{++} and dNTP bias, which prefers A:T (7). Libraries were cloned into a pUC-based plasmid and transformed into MV1932 for genetic complementation. Mutants from each library were sequenced before methyl methanesulfonate (MMS) selection. For detailed PCR mutagenesis and cloning methods, refer to *Supporting Methods* and Tables 4–6, which are published as supporting information on the PNAS web site.

Genetic Selection for Active Enzymes. MV1932 cells were transformed with pGRFP2-AAG, low, medium, and high libraries and with empty pGRFP2 vector and grown to confluence, diluted 1:100, and grown to midlogarithmic phase in LB-carbenicillin (LB-carb) at 37°C. Cultures were treated with 0.2% MMS for 1 hr, and the drug was washed away. Pretreated and posttreated cultures were serially diluted and plated on LB-carb in triplicate to calculate survival means and standard deviations. The fractions of surviving clones in libraries were normalized to wild-type survival. The dose of MMS used was within a range of drug in which the library and control populations were proportionally affected. MMS sensitivity assays were also performed at 0.15% and 0.25% MMS; and the percent library survivals relative to controls and each other were similar at these MMS concentrations (data not shown).

Supporting Methods present methods used for (*i*) PCR mutagenesis, (*ii*) DNA sequencing, (*iii*) AAG protein activity assay, and (*iv*) *x* factor calculation and protein substitutability visualization.

Results and Discussion

Calculating Protein Tolerance to Random Amino Acid Substitutions. The probability of protein inactivation with one random amino acid substitution, the *x* factor (x_{sub}), can be calculated from the

BIOCHEMISTRY

Abbreviations: AAG, human 3-methyladenine DNA glycosylase; MMS, methyl methanesulfonate.

‡To whom correspondence should be addressed. E-mail: laloeb@u.washington.edu.

Table 1. Calculating the *x* factor

	% of library with (*n*) number of amino acid changes ($f_n \times 100$)													Average mutation frequency	Library size	% Indels ($i \times 100$)	% Survival ($S \times 100$)	*x* factor (x_{sub})
	0	1	2	3	4	5	6	7	8	9	10	11	12					
WT-AAG	100	—	—	—	—	—	—	—	—	—	—	—	—				100 ± 8.6	
Low	5	20	40	20	15	0	0	0	0	0	0	0	0	2.2	2×10^5	6.1	32.7 ± 3.5	0.39 ± 0.038
Medium	5.6	5.6	5.6	11.1	5.6	33.3	22.2	5.6	0	5.6	0	0	0	4.6	1×10^5	9.9	18.2 ± 3.3	0.30 ± 0.052
High	0	3.6	7.1	10.7	3.6	10.7	14.3	10.7	17.9	14.3	3.6	3.6	0	6.2	0.9×10^5	5.5	10.7 ± 2.3	0.33 ± 0.034
Vector only	—	—	—	—	—	—	—	—	—	—	—	—	—				0.0051 ± 0.00017	
																	Average *x* factor	0.34 ± 0.06

Distribution of amino acid mutation load frequencies (f_n), library survival (S), and x factors (x_{sub}) in the low, medium, and highly mutated libraries. Indels (i) are expected to produce nearly 100% inactivation and are thus subtracted from the unadjusted x factors (x_T) to yield x factors due to amino acid substitutions (x_{sub}). As expected, increasing average mutation load results in lower percentage of active enzymes.

fractions of mutants (amino acid mutation load frequencies, f_n) with (n) number of amino acid changes within a gene-wide randomly mutated library, and from the proportion of mutants that survive functional selection (S). For example, f_0 denotes the fraction of the unselected library with 0-aa change, f_1 denotes the fraction with 1-aa change, and so on:

$$f_0(1-x_T)^0 + f_1(1-x_T)^1 + f_2(1-x_T)^2 + f_n(1-x_T)^n + \cdots = S \text{ or } \sum_{n=0} f_n(1-x_T)^n = S, \quad [1]$$

where x_T is the total protein inactivation probability with random amino acid change, including frameshifts (indels, i). x_T can be solved after experimental determination of the f_n, S, and i values. Indels are found at low percentages in random mutagenesis libraries, but invariably produce protein inactivation. To determine the true x factor (x_{sub}) resulting only from a random codon substitution (missense or nonsense mutation), the indel fraction in the total mutational pool (i) is subtracted from x_T to obtain x_{sub}.

$$X_{sub} = X_T - i \quad [2]$$

To measure the probability of inactivation by random amino acid substitutions, we used the gene encoding the human AAG. AAG protects cells against DNA alkylation damage by excising alkylated base lesions including 3-methyladenine, 7-methylguanine, and 1,N^6-ethenoadenine (εA) (8). The 894-bp AAG cDNA encodes a 298-aa 33-kDa monomeric protein that complements the DNA alkylation repair-deficient strain MV1932 (*ada alkA1*) (6) against toxicity induced by the alkylating drug MMS (9). Under MMS challenge, MV1932 cells expressing AAG from our pUC based vector exhibit >19,000-fold survival advantage over non-AAG-expressing MV1932 controls (Table 1), thus providing a stringent and specific selection for active mutant AAG enzymes.

The crystal structure of the catalytically competent ΔN79 (residues 80–298) AAG protein complexed with an 1,N^6-ethenoadenine substrate oligo reveals that the enzyme binds to DNA via a flat positively charged face. A β-hairpin extends into the DNA minor groove and flips the targeted nucleotide into the enzyme active site (10, 11). A water molecule is deprotonated by Glu-125 to form a hydroxyl nucleophile that cleaves the glycosylic bond between the damaged base and the sugar. The resulting abasic site is later cleaved and replaced with a normal nucleotide by the subsequent actions of an endonuclease, a DNA polymerase, and a DNA ligase (8).

We used PCR mutagenesis to generate low, medium, and highly mutated AAG cDNA libraries averaging 2.2-, 4.6-, and 6.2-aa changes per gene (a change is defined as a missense, nonsense, or indel). Sequencing of 20, 18, and 28 mutants from each unselected library revealed the f_n and i values of each library (Table 1). Expression of AAG and AAG mutant libraries protected MV1932 cells against MMS-induced cell death. The fractional survival of each library relative to wild-type yielded the S values (Table 1). Solving for x_{sub} using Eqs. 1 and 2 yielded the x factors (x_{sub}) of the low, medium, and high libraries at: 39% ± 4%, 30% ± 5%, and 33% ± 3% (mean ± standard deviation), respectively. The x factors from the three libraries are within the 95% confidence interval of each other. The average x factor is 34% ± 6%. Thus, the overall probability of inactivating AAG with a single random amino acid change occurring randomly in the protein is ≈34%, or one-third (Table 1).

The *x* Factor and the Substitutability of Proteins. Using three different libraries, we obtained a consistent value for the probability that a random amino acid change will inactivate AAG. Our findings beg the question of whether a similar x factor is seen in other proteins. It may be argued that the wide range of protein functions should demand drastically different mutabilities of various proteins. On the other hand, proteins face essentially similar requirements, such as the need to properly fold into soluble globular structures necessary for function (12). General types of changes leading to unfolding would inactivate various proteins. To address these questions, we reanalyzed data from diverse published studies and calculated inactivation probabilities. First, we examined random oligonucleotide mutagenesis studies in which mutations were targeted to the catalytic center of enzymes, and from which f_n and S data are available (13–18). We reasoned that these critical segments are expected to tolerate few substitutions. The results from human, bacterial, and viral enzymes are shown in Table 2. Despite the different enzymes and selection systems used, inactivation probabilities within these sensitive regions range from 44% to as high as 81%, averaging ≈60%, thus supporting our hypothesis. Second, Markiewicz and coworkers (19) examined 12 or 13 different amino acid substitutions at each residue across 90% of the 360-aa *E. coli lac* repressor protein using amber codon suppressor strains, which often corresponded to two or three nucleotide changes per codon. In our reanalysis of their data, we counted close to 1,380 single mutants that were inactive, ≈20% of which were temperature sensitive, of a total of 4,049 examined. This yielded an x factor for the *lac* repressor gene of 34%, which correlates well with our results for human AAG. Third, the x factor of a protein is conceptually similar to the proportion of new deleterious alleles that arise during the evolution of the source organism. Eyre-Walker and Keightley (20) calculated the percentage of deleterious substitution mutations that were eliminated from the human lineage by purifying selection. They examined synonymous and nonsynonymous substitution rates from coding regions of 46 homologous proteins from humans and chimpanzees. Interestingly, they conclude that at least 38% of spontaneous mutations in the human lineage were sufficiently deleterious to have been eliminated by selection (20). Together, these findings from multiple and independent experimental ap-

Table 2. x values calculated from active-site targeted cassette mutagenesis studies

Protein	Organism	Protein region (amino acid no.)	Average substitution frequency	Survival fraction (S)	x value	Ref.
DNA polymerase η	*Homo sapiens*	52–73	3.8	0.02	0.81	13
DNA polymerase I	*Thermus aquaticus*	605–617 (Motif A)	3.8	0.04	0.59	14
Thymidylate synthase	*H. sapiens*	196–199, 204–212	4.2	0.1	0.6	15
Reverse transcriptase	HIV	67–78 (β3 β4 loop)	4.1	0.11	0.59	16
DNA polymerase I	*Thermus aquaticus*	659–671 (O helix)	2.7	0.11	0.8	17
Thymidine kinase	*Herpes Simplex-1*	155, 161–165	2.4	0.32	0.44	18

Available f_n and S values were used to derive the inactivation probabilities. Although the various complementation systems may require differing levels of minimal enzyme activity, nevertheless the x values were >34% due to the concentration of mutations near the enzyme active sites.

proaches suggest a range of similar x factors over the length of diverse proteins.

Enzyme inactivation can result from indirect structure disrupting mutations or from direct alterations of the catalytic mechanism. The AAG functional assay is sensitive to both modalities. Minimal AAG activity necessary for complementation was assessed by measuring initial reaction rates under saturating substrate conditions in lysates of 10 random surviving clones. The results indicated that ≈5–10% of wild-type activity is necessary for survival at the MMS dose used (data not shown). Hydrophobic/hydrophilic properties appear to be crucial overall determinants of protein structure (5, 12). The buried core is sensitive to nonhydrophobic changes and those that disrupt packing, whereas the solvent-accessible surface is generally more tolerant of change. Residue size, charge, hydrogen-bonding characteristics, and bond angle flexibility are other folding factors that may be perturbed by random substitutions.

AAG is a simple monomeric protein. Larger proteins with multiple functional domains and multiple interacting partners may exhibit more complex inactivation dynamics. The x factor calculation assumes that the effects of multiple mutations are independent, in that the effects of mutations on protein function are largely additive. This is supported by findings on the λ repressor (21). However, at higher mutational loads effects of mutations may interact in more complex ways, with increased possibility of compensatory or synergistic effects. These results with AAG may slightly underestimate the x factor, because the N-terminal 79 aa of AAG are not required for enzymatic activity. Tolerated substitutions are slightly elevated in this N-terminal region. Nevertheless, protein-folding principles apply, and mutations in this region that cause overall misfolding or aggregation will produce inactivation.

There likely are variations in the substitutability of different proteins. The hydrophobic core is generally less tolerant of change than the solvent accessible exterior (5). Therefore, x factors may also be influenced by protein sizes and surface-to-volume ratios. Axe and coworkers found that 5% of single amino acid substitutions lead to an inactivated barnase enzyme (22). Rennell *et al.* (23) found that ≈16% of amino acid substitutions in T4 lysozyme caused inactivation. The differences from the above findings may be attributed to barnase and T4 lysozyme small sizes, which are 110 and 164 aa, respectively. Highly conserved proteins such as histones are likely to be relatively intolerant to mutation, whereas protein domains such as F_v regions of antibodies may exhibit increased tolerance against misfolding. Residues that are posttranslationally modified are also expected to be intolerant of change.

The x factor is calculated for amino acid replacements and can include the generation of stop codons. The frequencies of stop codons in the low, medium, and high libraries are 4%, 9%, and 7.5%, respectively. The x factor can be converted for single-nucleotide substitutions. Largely due to degeneracy at the third position, the nucleotide x factor is expected to be less than the amino acid x factor. Multiplying the amino acid x factor of ≈34% by the probability of nonsynonymous codon change accessible by one nucleotide (415/549) yields the nucleotide x factor of ≈26%.

Our mutagenesis scheme of creating predominantly random single nucleotide substitutions mimics the generation of natural diversity. Three naturally occurring human single-nucleotide polymorphisms arose in our database of tolerated AAG substitutions: P64L, T199A, and A258V. These variations did not exhibit appreciable effects on MV1932 complementation when individually assayed (data not shown).

Substitutability and Structure. Previously, we have focused on the probability of amino acid changes being inactivating. We have also examined situations in which amino acid substitutions are tolerated. To analyze the nature of tolerated substitutions, we sequenced 244 mutant AAG cDNAs from the highly mutated library that complemented MV1932. This yielded a total of 920 tolerated amino acid changes. Fig. 1 maps the mutations along the AAG primary sequence. The types of tolerated amino acid substitutions at each position are indicated. Residues without bars reflect zero identified substitutions.

A residue's "substitutability index" is defined as the percent sequenced clones with a substitution at that residue. Many positions that are evolutionarily conserved are also essential for activity (10, 11) and did not tolerate changes in our assay. Examples include Glu-125, Arg-182, and Val-262, each of which interacts with the activated water molecule that hydrolyzes the sugar-base glycosylic bond. Other nonsubstituted amino acid residues include Tyr-162, which projects from a surface β hairpin and acts as a "nucleotide flipper." Met-164 and Tyr-165 assist in this base-flipping mechanism by destabilizing the base pair adjacent to the flipped nucleotide. Y162A, M164A, and Y165A single substitution mutants were generated by Lau *et al.* (11) and assayed by using a genetic complementation system. The Y162A mutant exhibited large impairment of glycosylase activity, whereas M164A and Y165A showed only moderate impairment (11). Correspondingly, in our study, no substitutions were observed at Tyr-162, whereas positions Met-164 and Tyr-165 showed moderate substitutability, allowing Ile, Arg, and Phe substitutions, respectively (Fig. 1). Within the substrate-binding pocket, the flipped-out base stacks between the aromatic side-chains of Tyr-127, His-136, and Tyr-159. Y127F, H136Q, and Y159F mutants were also generated previously (11). Y127F exhibited the most profound decrease in activity, whereas Y159F was the least affected (11). In our data set, Tyr-127 was concordantly unsubstituted, and His-136 tolerated only one Tyr replacement. Tyr-159 was substituted by both Phe and Asn.

There are positions in AAG that are not evolutionarily conserved but did not exhibit any tolerated changes. The individual spatial arrangements of these interactions are likely unique to AAG.

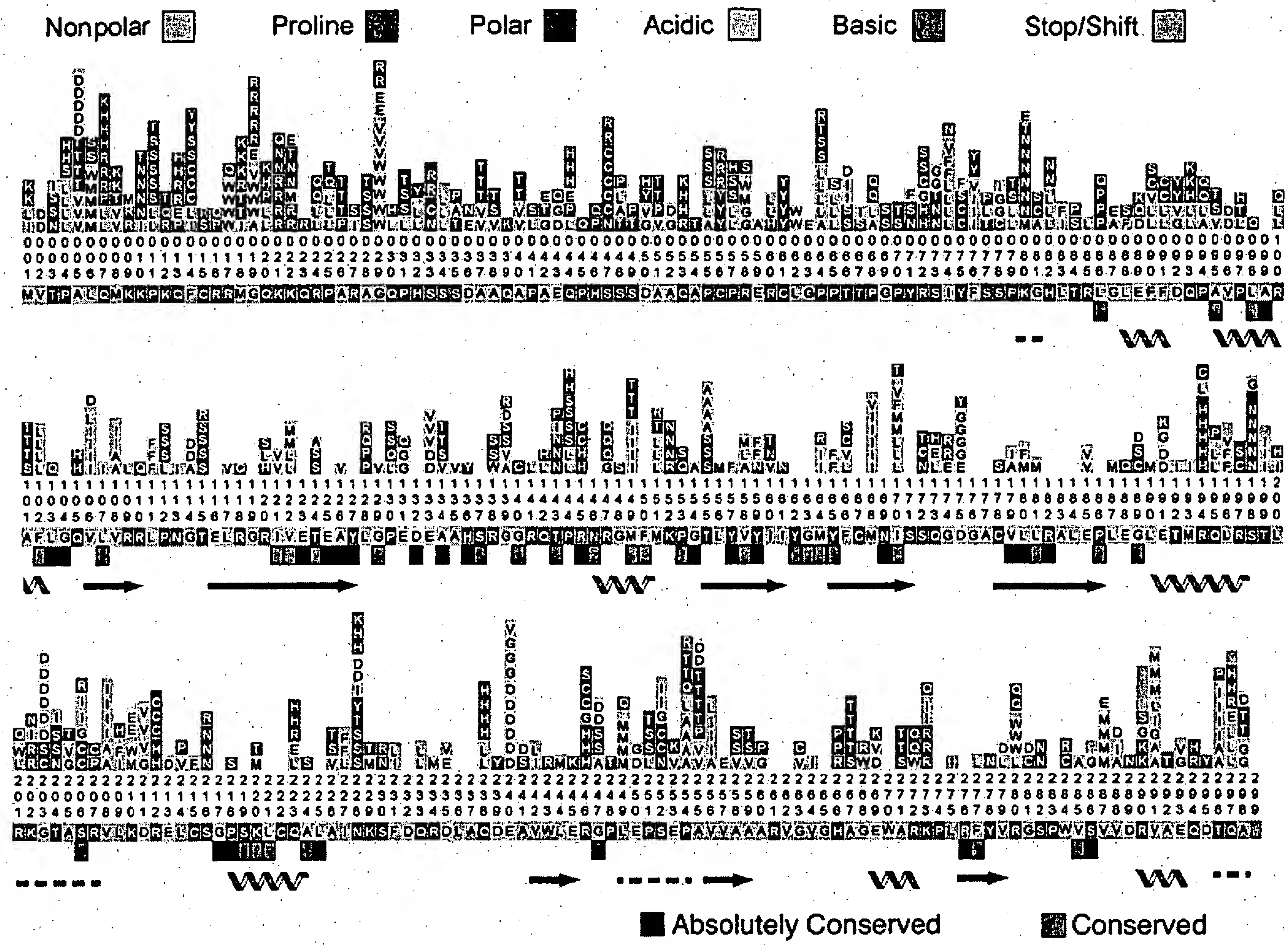


Fig. 1. Tolerated amino acid changes along the AAG primary sequence, shown with evolutionary conservation and secondary structures. Two hundred forty-four active AAG mutants were sequenced, and observed amino acid substitutions are shown above the wild-type sequence. Colored bars indicate general categories of amino acids. Numbers are read vertically and indicate residue position. Below the wild-type sequence, evolutionarily invariant residues are marked by black boxes and conserved residues by gray boxes. α helices (helix), β strands (arrows), and disordered regions (dashed lines) are indicated. Homologous sequences (from human, mouse, rat, *Borrelia burgdorferi*, *Bacillus subtilis*, *Arabidopsis thaliana*, and *Mycobacterium tuberculosis*) were identified with PSI-BLAST (10), and secondary structure calling was performed with MOLECULAR OPERATING ENVIRONMENT (MOE, CCG, Montreal, Canada).

Although some of these positions may display substitutions if even more mutants are sequenced, the structural basis for lack of substitutions at many of these positions highlights three general mechanisms: specific hydrogen-bonding interactions, unique hydrophobic packing, and ion binding. For example, specific hydrogen-bonding requirements are emphasized by Glu-116's interaction with Arg-118, which, in turn, interacts with Glu-188 and Glu-245 in a three-way interaction. Arg-261 provides a hydrogen pair partner to the evolutionary conserved and unsubstituted Asp-132. This pair packs adjacent to Tyr-127, which forms part of the active site pocket. Hydrophobic packing constraints are observed at Gly-119, which is at the core of a β strand, <4.5 Å away from Leu-184. No other side chains can fit in this tight space. Similar packing constraints are observed at Leu-184, which is <4.5 Å from the unsubstituted Leu-225. Cys-167 is buried <4.5 Å from Ile-227 and close to the Cα of Cys-222. Mutations of buried residues may require concomitant mutations of other closely packed residues to maintain optimal packing. Interestingly, at least one mutant in our study appears to demonstrate this principle. It contains I170V and L181M substitutions that pack adjacently in the hydrophobic core. The conversion of Leu-181 to the slightly bulkier methionine is found to coexist with the conversion of Ile-170 to the smaller valine. Last, lack of substitutions at Ser-171 highlights the role of ion binding. Ser-171's side-chain oxygen binds to a Na^+ ion, which has been postulated to enhance the structural stability of the active site floor (11).

In contrast, certain regions in AAG appear highly substitutable. Examples include the first 79 N-terminal residues that have been shown previously to be unnecessary for *in vitro* enzyme activity and DNA-binding specificity (24). Residues 80–81, 200–207, 249–254, and 296–298 are also highly substitutable (Fig. 1). In accord, they display low electron density in x-ray crystallography and were inferred to be disordered loops (10).

In Fig. 2, the relative substitutability indices of residues are mapped onto the available crystal structures of the N79Δ AAG mutant. Dark-blue residues are the least substitutable, and red residues are the most tolerant of change. Fig. 2 *A* and *B* shows surface residues, and Fig. 2 *C* and *D* facilitates views into the protein core. One striking feature is the general immutability of the DNA-interacting face and specifically, the nucleotide-flipper Tyr-162 (Fig. 2*A*). A surface region distant from the DNA-binding face (Fig. 2*B*) was also observed to have low substitutability scores; Glu-188, Arg-118, Glu-245, Glu-116, and Arg-110 participate in a network of charged contacts that likely contribute to protein

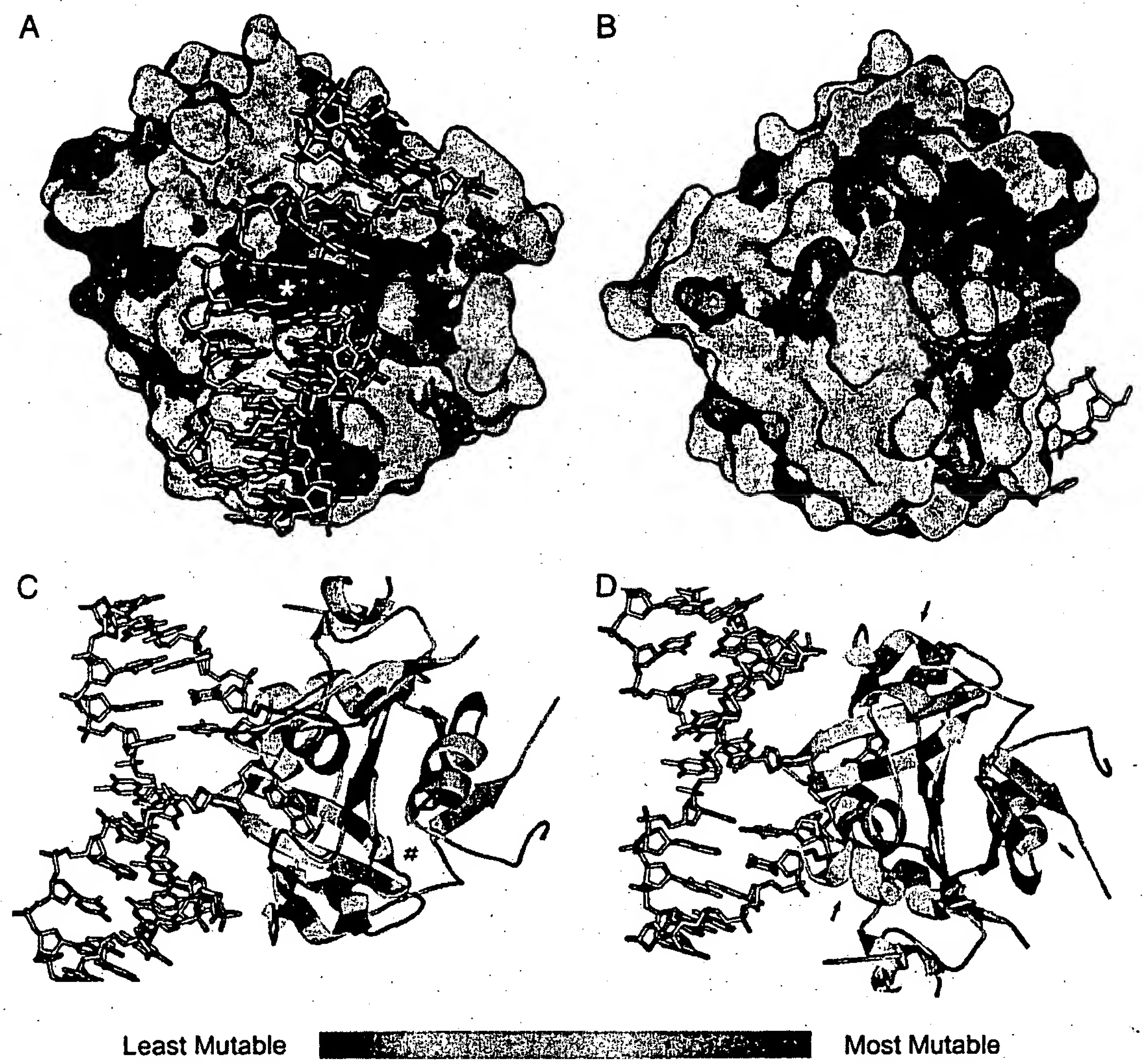


Fig. 2. Substitutability of AAG amino acid residues and structure. Individual residues' substitutability scores are indicated by their color in the spectrum, with red being the most substitutable and dark blue the least. (*A*) The DNA interacting face of AAG. The DNA-binding face and intercalating Tyr-162 (*) are largely intolerant of substitution, whereas distant loops are generally tolerant of change. (*B*) Rotation of *A* by 180°, showing the opposite side of the AAG surface. (*C* and *D*) The substitutability of AAG resides shown by secondary and tertiary structure representation. Views are rotated 180° relative to each other. Residues near the active site of AAG, adjacent to the extrahelical and 1,N^6-ethenoadenine DNA lesion, are generally intolerant of change. The β4 (165–171) strand is indicated by #. Arrows point toward α helices with solvent accessible faces that exhibit greater substitutability than their buried sides.

stability. In the protein interior, a conspicuous pattern of alternating unsubstituted and substitutable sites is seen in the β4 (165–171) strand (Figs. 1 and 2 *C* and *D*). Cys-167, Asn-169, and Ser-171 are relatively unsubstituted, because their side chains face toward the active site and are involved in substrate recognition or Na^+ binding (11). In contrast, Met-168 and Ile-170 tolerate hydrophobic substitutions, because their side chains face the opposite direction and pack into the hydrophobic core. Solvent-accessible surfaces generally exhibit higher substitutability compared with buried residues. This is evident in Fig. 2 *C* and *D*, where the exposed exterior sides of several α helices exhibit greater substitutability than their interior-facing sides.

Averages of substitutability indices in different structural motifs are presented in Table 3. In AAG, evolutionarily conserved and catalytically crucial residues are significantly less substitutable than the rest of the protein. Nonconserved residues adjacent to conserved residues in the primary sequence are generally less substitutable than other nonconserved residues, reflecting their involvement in functionally important regions. This observation suggests they may also be fruitful targets for directed evolution studies. β strand residues, as a group, are less tolerant of substitution than are α helices. This may be explained in part by the fact that in this α-β protein, the β-sheets are generally less solvent accessible and therefore possess fewer surface residues that are more likely to tolerate substitutions. Loops and turns, expectedly, are the most substitutable.

Some Implications of the *x* Factor. We observed that various residues of a protein are differentially sensitive to substitutions, and that tolerance of the entire protein to random change can be defined by the *x* factor. The *x* factor is a description of an intrinsic property of individual proteins and protein motifs and can be a guiding parameter in the study of natural and artificial evolutionary processes. For example, using the estimated inactivation probability of ≈34% and assuming mutually independent effects on inactivation probability by multiple mutations, the isolation of active mutants harboring many mutations from large random mutagenesis libraries ($>10^5$) is not surprising (25). In contrast, a single, non-3-bp indel event almost certainly leads to inactivation ($x \approx 1$). Therefore, indel frequencies should be minimized

Table 3. Mean substitutability indices of AAG motifs

	Motif residue substitutability	Nonmotif residues	*t* test
	Mean ± SD	Mean ± SD	*P* value
Entire protein	1.38 ± 1.10		
Evolutionarily conserved	0.75 ± 0.85	1.53 ± 1.10	6.37E-08
Nonconserved adjacent to conserved	1.25 ± 0.84	1.60 ± 1.16	0.017
α-helices	1.19 ± 0.97	1.41 ± 1.12	0.19
β-strands	0.73 ± 0.76	1.52 ± 1.12	1.01E-08
Turns and loops	1.57 ± 1.12	0.93 ± 0.89	3.46E-07
Functionally important residues	0.56 ± 0.60	1.41 ± 1.11	1.04E-04

The substitutability index of individual residues (number of observed amino acid changes/number of active mutants sequenced at that position) is expressed as a percentage (×100), categorized by motifs, and averaged. The *t* test is performed against indices of motif nonmembers for differences in mean substitutability indices, indicative of differing importance to enzyme function.

in efforts to evolve novel proteins from high mutation load libraries. Retroviruses, such as HIV, may be susceptible to increased mutational burden, and lethal mutagenesis of viral genomes by introducing mutations through the use of nucleoside and ribonucleotide analogs has been proposed (26). Given our findings, such efforts may be further enhanced by the use of analogs that efficiently induce frameshift mutations. Viral genomes that encode multiple proteins as different reading frames of the same genetic sequence may be particularly sensitive to agents that generate frameshifts.

It is estimated that the human mutation rate per coding diploid genome per generation is 3.2, including base substitutions, indels, and larger changes (27). Multiplying this number by the general *x* factor of ≈34%, the rate of introducing deleterious coding alleles by random substitution is ≈1.0 per diploid genome per sexual generation. This is likely an underestimate, because indels inactivate coding regions much more efficiently than base substitution mutations. Dominant negative mutations may also more efficiently produce a deleterious phenotype, although the frequency of mutations that act in a dominant negative manner is largely unknown. Interestingly, our deleterious coding allele rate calculation of 1.0 is congruent with the estimate of 1.6 independently calculated by Eyre-Walker and Keightley (20), which was based on the assumption of 60,000 genes in the human genome.

Overall, our method of gene-wide random mutagenesis and sequencing highlights the relative importance of specific residues to enzyme structure and function through the numbers and types of tolerated substitutions. This work validates and extends from previous structural studies. Interestingly, the substitutability indices of individual residues can be obtained independently of conservation or structural information and are generally consistent with both. The extensive database of tolerated amino acid substitutions is obtained from a more expedient form of gene-wide study than previous techniques, such as alanine scanning. This database can provide a valuable resource for predicting the effects of mutations on protein function, which has been a focus of recent investigations (28, 29).

We advance the concept of the *x* factor as a measure of protein tolerance to random substitutions. The *x* factor may also be useful in measuring genomic robustness against mutations. It has been hypothesized that evolvability, or the ability to generate heritable variation, may be favored in certain environments (30). Genomes experiencing high mutational burden may face selective pressure to evolve proteins that are tolerant of change, in which case the observed *x* factors are expected to be less than *x* factors of homologous proteins from more faithfully propagated genomes. It may be of particular interest to examine *x* factors from various protein families and diverse organisms.

We are indebted to Dr. Elinor Adman for structural discussions and modeling; Drs. Greg Ireton and Django Sussman for help in generating Fig. 2; Drs. Steve Henikoff, David Baker, and Michael Fry for discussions; Drs. John Davidson, Ann Blank, and Raymond Monnat for critical reading of the manuscript; and Dan Shen for X-CALCULATOR web site programming. This work was supported by National Institutes of Health Grants CA78885 and CA80993. H.H.G. and J.C. are also supported by the Medical Scientist Training Program of the University of Washington.

1. Loeb, L. A., Loeb, K. R. & Anderson, J. P. (2003) *Proc. Natl. Acad. Sci. USA* **100**, 776–781.
2. Harris, R. S., Sheehy, A. M., Craig, H. M., Malim, M. H. & Neuberger, M. S. (2003) *Nat. Immunol.* **4**, 641–643.
3. Smith, J. M. (1970) *Nature* **225**, 563–564.
4. Creighton, T. E. (1993) *Proteins* (Freeman, New York).
5. Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A. & Sauer, R. T. (1990) *Science* **247**, 1306–1310.
6. Posnick, L. M. & Samson, L. D. (1999) *J. Bacteriol.* **181**, 6763–6771.
7. Kawate, H., Landis, D. M. & Loeb, L. A. (2002) *J. Biol. Chem.* **277**, 36304–36311.
8. Wyatt, M. D., Allan, J. M., Lau, A. Y., Ellenberger, T. E. & Samson, L. D. (1999) *BioEssays* **21**, 668–676.
9. Samson, L., Derfler, B., Boosalis, M. & Call, K. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 9127–9131.
10. Lau, A. Y., Scharer, O. D., Samson, L., Verdine, G. L. & Ellenberger, T. (1998) *Cell* **95**, 249–258.
11. Lau, A. Y., Wyatt, M. D., Glassner, B. J., Samson, L. D. & Ellenberger, T. (2000) *Proc. Natl. Acad. Sci. USA* **97**, 13573–13578.
12. Beasley, J. R. & Hecht, M. H. (1997) *J. Biol. Chem.* **272**, 2031–2034.
13. Glick, E., Vigna, K. L. & Loeb, L. A. (2001) *EMBO J.* **20**, 7303–7312.
14. Patel, P. H. & Loeb, L. A. (2000) *Proc. Natl. Acad. Sci. USA* **97**, 5095–5100.
15. Landis, D. M. & Loeb, L. A. (1998) *J. Biol. Chem.* **273**, 25809–25817.
16. Kim, B., Hathaway, T. R. & Loeb, L. A. (1996) *J. Biol. Chem.* **271**, 4872–4878.
17. Suzuki, M., Baskin, D., Hood, L. & Loeb, L. A. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 9670–9675.
18. Black, M. E. & Loeb, L. A. (1993) *Biochemistry* **32**, 11618–11626.
19. Markiewicz, P., Kleina, L. G., Cruz, C., Ehret, S. & Miller, J. H. (1994) *J. Mol. Biol.* **240**, 421–433.
20. Eyre-Walker, A. & Keightley, P. D. (1999) *Nature* **397**, 344–347.
21. Gregoret, L. M. & Sauer, R. T. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 4246–4250.
22. Axe, D. D., Foster, N. W. & Fersht, A. R. (1998) *Biochemistry* **37**, 7157–7166.
23. Rennell, D., Bouvier, S. E., Hardy, L. W. & Poteete, A. R. (1991) *J. Mol. Biol.* **222**, 67–88.
24. O'Connor, T. R. (1993) *Nucleic Acids Res.* **21**, 5561–5569.
25. Zaccolo, M. & Gherardi, E. (1999) *J. Mol. Biol.* **285**, 775–783.
26. Loeb, L. A., Essigmann, J. M., Kazazi, F., Zhang, J., Rose, K. D. & Mullins, J. I. (1999) *Proc. Natl. Acad. Sci. USA* **96**, 1492–1497.
27. Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. (1998) *Genetics* **148**, 1667–1686.
28. Ng, P. C. & Henikoff, S. (2001) *Genome Res.* **11**, 863–874.
29. Saunders, C. T. & Baker, D. (2002) *J. Mol. Biol.* **322**, 891–901.
30. Radman, M., Matic, I. & Taddei, F. (1999) *Ann. N. Y. Acad Sci.* **870**, 146–155.

ELSEVIER

Analytica Chimica Acta 391 (1999) 345–351

ANALYTICA CHIMICA ACTA

A colorimetric assay suitable for screening epoxide hydrolase activity

Frank Zocher[a], Markus M. Enzelberger[a], Uwe T. Bornscheuer[a], Bernhard Hauer[b], Rolf D. Schmid[a,*]

[a]*Institute for Technical Biochemistry, Stuttgart University, Allmandring 31, D-70569 Stuttgart, Germany*
[b]*BASF AG, Biotechnology Department, D-67056 Ludwigshafen, Germany*

Received 26 February 1999; accepted 28 February 1999

Abstract

A UV/VIS spectrophotometric microtiter-plate and a filter-paper based assay using 4-(*p*-nitrobenzyl)pyridine (NBP) were developed to determine epoxide hydrolytic activity by measuring the decrease of the epoxide concentration. Both systems were applied for screening an expression gene bank of *Rhodococcus* sp. NCIMB 11216. As a reference, whole cells from *Rhodococcus* sp. NCIMB 11216 and *Beauveria sulfurescens* ATCC 7159 exhibiting epoxide hydrolase activity were used. The microtiter-plate system was also evaluated for different epoxides and performed in a laboratory robotic system for high throughput screening. The microtiter-plate assay showed a high sensitivity for the detection of small concentrations of epoxides (0.1–1 mg/well) such as styrene oxide, ethyl phenylglycidate, *n*-hexane oxide and indene oxide. The filter paper assay was further optimized for styrene oxide. Both assays were suitable to screen within libraries of epoxide hydrolases without interference with other enzymes such as esterases, lipases or proteases. The assay should allow to screen large libraries obtained by directed evolution, strain collections and (expression) gene banks for epoxide hydrolytic activity or to monitor the purification process of an epoxide hydrolase.

Keywords: Colorimetric assay; Epoxide hydrolase; 4-(*p*-nitrobenzyl)pyridine; High throughput screening

1. Introduction

Epoxides are useful chiral compounds for organic chemistry and used as valuable intermediates for the synthesis of enantiomerically pure pharmaceutically active substances. They can be prepared by several chemical methods such as the Sharpless epoxidation of allylic alcohols [1] or by the method of Jacobsen and Katsuki [2]. They can also be obtained enantiomerically pure by microbial reduction of α-haloketones, by microbial degradation of halohydrins or by monooxygenase reactions catalyzed by different P450 species [3,4]. These epoxides are often cleaved by microsomal or cytosolic epoxide hydrolases which play an important role in the detoxification of xenobiotics [5]. Epoxide hydrolases (EC 3.3.2.3) belong to the class of α/β-hydrolase-fold enzymes. The reaction mechanism displays similarities to dehalogenases [6,7]. Different microbial and fungal epoxide hydrolase are known from literature, e.g. from *Rhodococcus* sp. NCIMB 11216 [8], *Aspergillus niger* LCP 521 [9], *Beauveria sulfurescens* ATCC 7159 [10] and *Nocardia* EH1 [11]. Recently, the first recombinant bacterial

*Corresponding author. Tel.: +49-711-685-3193; fax: +49-711-685-3196; e-mail: itbrsc@po.uni-stuttgart.de

0003-2670/99/$ – see front matter
PII: S0003-2670(99)00216-0

epoxide hydrolases were cloned from *Agrobacterium radiobacter* AD1 [12] as well as from *Corynebacterium* sp. C 12 [13] and the former was expressed in *E. coli*.

To determine epoxide hydrolase activity and enantioselectivity in hydrolysis, different methods were used such as liquid chromatography (LC) [14], gas chromatography (GC) [10] or UV/VIS [15]. LC and GC methods are not suitable to screen large mutant enzyme libraries obtained by directed evolution via error-prone polymerase chain reaction (PCR) [16] or gene shuffling [17], within strain collections or (expression) gene banks due to complicated sample preparation and time consuming analysis. UV/VIS spectrophotometric methods are more convenient because of simple and fast measurement which enables screening in high throughput systems. For filter paper assays, a chromophore indicating the hydrolysis of an epoxide in the visible area would be most suitable. Current assays determine the hydrolysis of epoxides by UV measurement and only the decrease of absorption is monitored. Unfortunately, the absorption coefficients are usually small [18,19]. Other assay substances such as epoxy carbonates, which release chromophores like *p*-nitrophenol are only suitable in the absence of other hydrolases such as esterases or lipases, which can cleave the carbonate. As a consequence, purified epoxide hydrolases are necessary [15].

For screening, engineering and for purification of new epoxide hydrolases, a fast and accurate assay is required which enables a high throughput. For the measurement of alkylating substances, the 4-(*p*-nitrobenzyl)pyridine (NBP) test is well known. It was applied to a variety of different alkylating substances such as reactive halogen compounds, carboxylic acid chlorides, organic phosphoric compounds, aziridines and epoxides [20]. The alkylation of NBP with styrene oxide analogs in vitro [21] and with chloroethylene oxide [22] were studied. A correlation of alkylating and mutagenic activities of allylic compounds was found using an NBP alkylating procedure [23]. Also, the mutagenic potential of slow-reacting epoxides was determined using NBP [24]. The NBP test was applied for the determination of microsomal epoxide hydrolase activity with safrole oxide (SAFO) as substrate after extraction/workup of the epoxide from the incubation medium [25]. A colorimetric assay and a thin layer assay using the NBP test after extraction of the reaction mixture with organic solvent and boiling of the assay mixture was developed [26].

In the present paper, we describe a modified NBP test using microtiter-plates together with a pipetting robot which can be performed directly with whole or lysed cells containing other hydrolytic enzymes without any extraction or separation steps at mild assay conditions in microtiter-plates or by using a filter-paper based assay for direct screening on agar plates.

2. Experimental

2.1. Reagents

All epoxides were obtained from Aldrich (Steinheim, Germany) at the highest purity available. Indene oxide was prepared using *N*-bromosuccinimide. The crude bromohydrin was treated with sodium bicarbonate according to literature [10] and the epoxide was purified by flash chromatography on silica gel (petroleum ether:ethyl acetate=9:1, R_f=0.55, yield 61%). ^{1}H- and ^{13}C-NMR spectra were in agreement with literature [10]. All other reagents were purchased from Fluka (Buchs, Switzerland) at the highest purity available.

2.2. Apparatus

UV/VIS spectrophotometric measurements were performed on a ICN Titertek MS 212 (ICN, Meckenheim, Germany) microtiter-plate reader at 560 nm (reference filter 650 nm) using standard 96 well microtiter-plates. Activity was calculated from the difference in absorbance at 560 and 650 nm.

GC analysis was performed on a FS-Cyclodex β-I/P CS-Fused Silica capillary column (CS-Chromatographie Service GmbH, Langerwehe, Germany) using H_2 as carrier gas (split 1:100, 130 kPa) at 75°C for styrene oxide and at 140°C for phenyl 1,2-ethandiol.

In case of the gene bank of *Rhodococcus* sp. NCIMB 11216, assays were done using a Biomek 2000 Workstation equipped with a Biomek SL side-loader with Beckman incubator and the software package BioWorks 2.2 from Beckman (München, Germany).

Normal Whatman filter paper (Springfield Mill, UK) was used for the agar plate assay.

2.3. Strains

Rhodococcus sp. NCIMB 11216 and *Beauveria sulfurescens* ATCC 7159 were cultivated in LB media (yeast extract 5 g/l, tryptone 10 g/l, NaCl 10 g/l, pH 7.0) at 30°C and 210 rpm in 2 l shaking flasks for 2 days. *E. coli* DH5α [F^- *endA1 hsdR17* ($r_k^- m_k^+$) *supE44 thi-1 recA1 gyrA96 relA1* Δ(*lac-ZYA-argF*)*U169* ϕ80d*lac*ZΔM15 λ^-] was cultivated in LB media at 30°C until $OD_{600\ nm}$ reached 0.35. Cells were harvested by centrifugation at 4000 rpm at 0°C. Cell lysis was performed by addition of 10 ml lysozyme solution (0.1 M Tris/HCl buffer, 300 mM NaCl, 1 mg/ml lysozyme) at 0°C for 1 h to the remaining cell pellet.

2.4. Recombinant DNA techniques

Rhodococcus sp. NCIMB 11216 cells were lysed by the addition of lysozyme as described above. Chromosomal DNA was partially digested by *Sau* 3A and ligated to vector pUC 19 that had been digested with *Bam* HI and dephoshorylated with alkaline phosphatase from calf intestine. *E. coli* DH5α was transformed with the ligation mixture and plated onto LB agar plates containing 100 µg/ml ampicillin. Transformants were subsequently replica-plated to these LB-ampicillin plates, which were used to screen for colonies with epoxide hydrolase activity. All operations were done using standard procedures [27]. Restriction enzymes, alkaline phosphatase and pUC 19 were obtained from MBI Fermentas (St. Leon-Rot, Germany).

2.5. Assay procedures

2.5.1. Assay in 96 well microtiter-plates

For validation, different amounts of styrene oxide (0.17 M), phenyl ethylglycidate (0.14 M), indene oxide (0.16 M), *n*-decane oxide (0.1 M) and *n*-hexane oxide (0.8 M) were dissolved in acetone and used as stock solutions. Stock solution (40 mg) was added to 100 µl of LB media. Different amounts of triethylamine (0, 18 or 36 mg) or piperidine (0, 22 or 43 mg) and 50 mg of triethylene glycol dimethylether were added. Then 50 mg of 4-(*p*-nitrobenzyl)pyridine solution (0.23 M NBP in methoxyethanol) was added and the reaction mixture was incubated at 39°C. The absorbance at 560 nm was measured at different times against 650 nm as reference wavelength and the difference in absorption was calculated.

Assay with E. coli DH5α cells: A stock solution of styrene oxide (1.3 M in acetone) was prepared and diluted with acetone (1.3, 0.65, 0.32, 0.16, and 0.08 M). These dilutions (40 mg) and 50 µl of whole or lysed *E. coli* DH5α cells were used and the assay was performed as described above.

2.5.2. Assay on filter paper

The filter papers were preincubated for 10 min in a glass petri dish containing a styrene oxide solution (90 mM in acetone). After air drying for 20 min, the agar plates were replica-plated and the colonies were transferred from the agar plates by manual-pressing the filter paper on the agar. The filter paper was incubated for 30 min at 37°C in a closed glass petri dish for epoxide hydrolysis. (Care should be taken due to the high toxicity of styrene oxide!) For developing, the filter paper was incubated in 2 ml of NBP solution (0.23 M NBP in methoxyethanol) for 1 min, air dried for 30 min and incubated for 30 min at 80°C in a closed glass petri dish. Hydrolysis activity could be monitored by the formation of a colorless zone on the blue filter paper.

3. Results and discussion

For the validation of the NBP assay in microtiter-plates, different epoxides were used (styrene oxide, phenyl ethylglycidate, indene oxide, *n*-hexane oxide and *n*-decane oxide). The assay showed a high sensitivity for all epoxides except for *n*-decane oxide because of the low alkylating strength of this epoxide. A linear correlation between the amount of epoxide and absorbance was found for all epoxides (data for *n*-hexane oxide not shown), with the exception of *n*-decane oxide (data not shown), in the microtiter-plate assay (Figs. 1–3). In the literature, a stabilizing effect by the addition of different bases especially piperidine to the assay mixture has been described [20]. In contrast, we observed that triethylamine or no addition of base gave better results. The addition of piperidine even led to a decrease in absorbance.

Although the fastest reaction in the NBP assay was observed for indene oxide compared to the other epoxides, the assay was further optimized for styrene oxide, because an easy evaluation of the results was

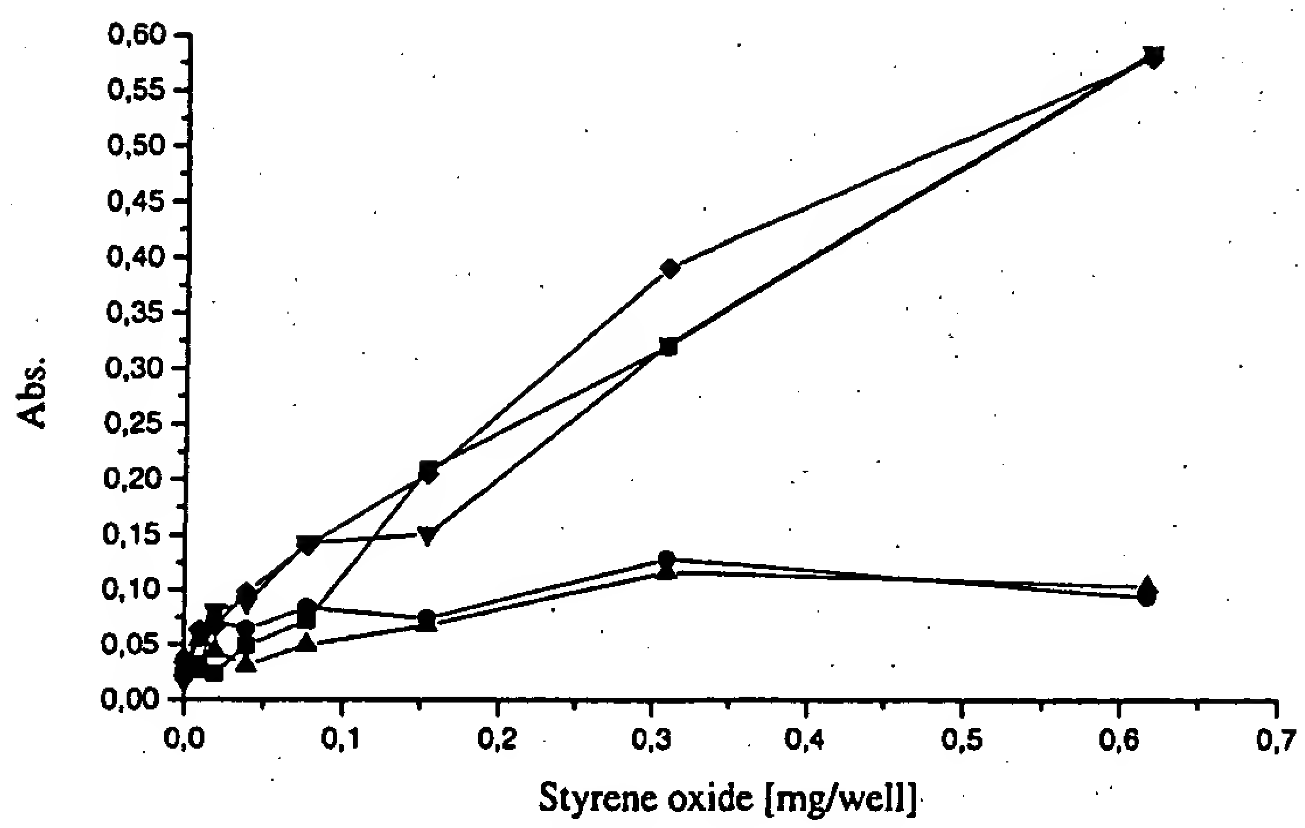


Fig. 1. Effect of incubation of styrene oxide on absorbance (60 min, 39°C) without base (■), with piperidine (25 μl ▲, 50 μl ●) and with triethylamine (25 μl ◆, 50 μl ▼).

possible by GC analysis. A clear correlation of the decrease in styrene oxide concentration as determined by the NBP assay with the consumption of styrene oxide and concomitant formation of the product phenyl 1,2-ethandiol as determined by GC analysis was observed.

The absorbance of reaction mixtures without and with different concentrations of styrene oxide was measured in the presence of *E. coli* DH5α cells in LB media. In the presence of *E. coli* cells, a linear correlation between the amount of styrene oxide and the absorbance could be obtained after 45 min (Fig. 4). The error introduced by measurement of the absorbance after 45 min is within ±3% for three control experiments under the same conditions. This means, that it is not necessary to do any extraction or work-up of the incubation mixture to remove the cells. The assay was also found to be accurate in the presence of lysed *E. coli* DH5α cells. Even in the presence of lysis products and remaining cell debris, a proportionality between the amount of styrene oxide and absorption was observed (Fig. 4).

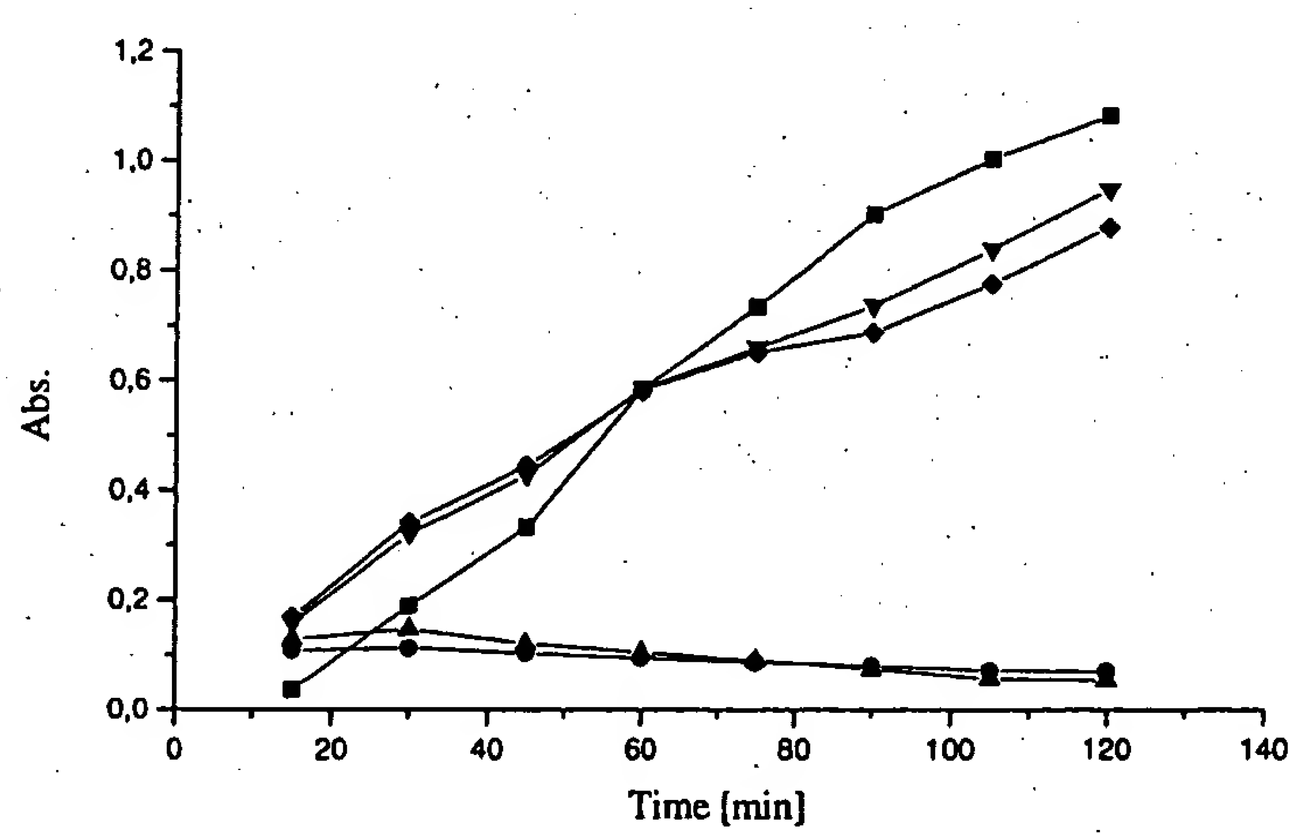


Fig. 2. Influence of reaction time on the hydrolysis of styrene oxide, and hence absorbance at different concentrations of bases (39°C, 0.62 mg styrene oxide/well) without base (■), with piperidine (25 μl ▲, 50 μl ●) and with triethylamine (25 μl ◆, 50 μl ▼).

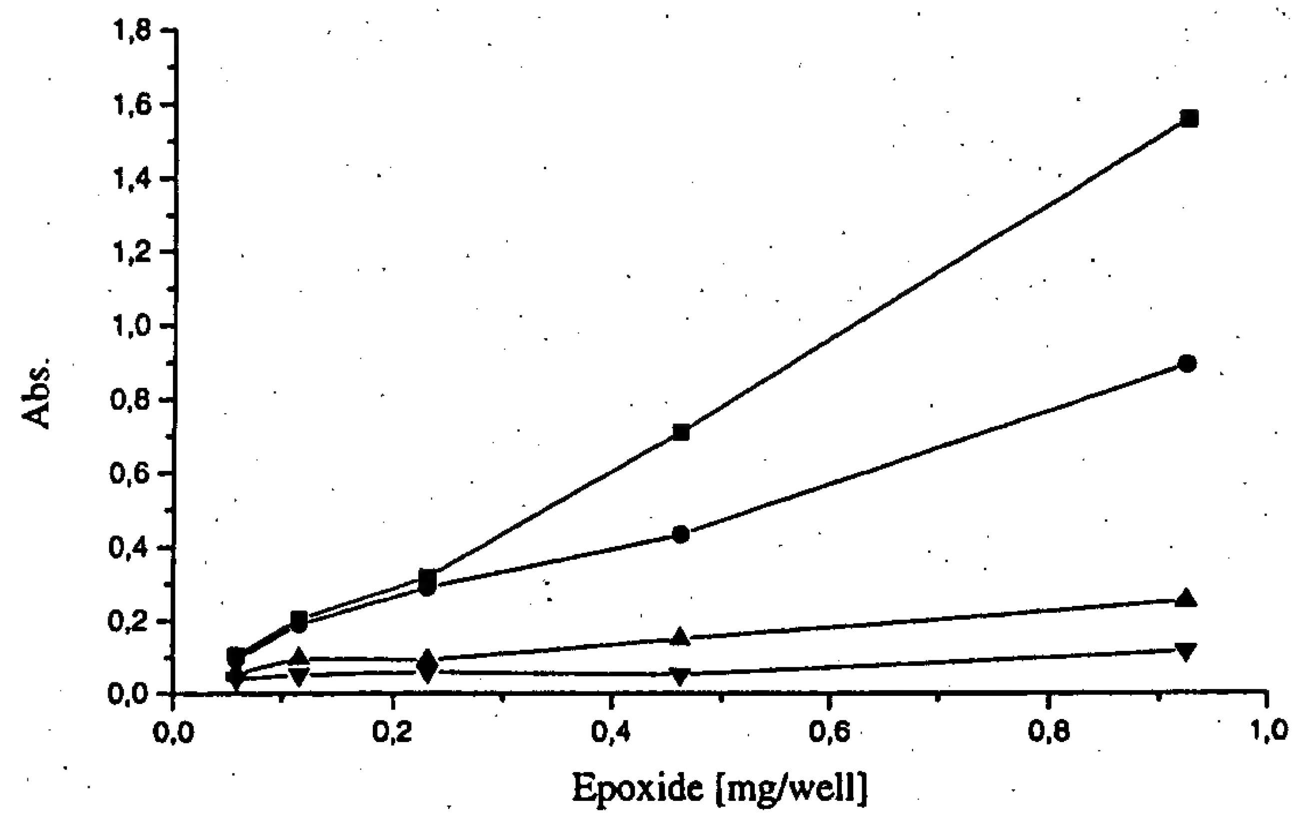


Fig. 3. Absorbance at different concentrations of indene oxide without base (■), with triethylamine (50 µl ●) and ethyl phenylglycidate without base (▲) and with triethylamine (25 µl ▼) (60 min incubation, 39°C).

Assays were also performed using 100 µl of a suspension of whole cells from *Rhodococcus* sp. NCIMB 11216 and *Beauveria sulfurescens* ATCC 7159 in LB media. After 2 h of incubation at 39°C in the presence of different dilutions of styrene oxide (same conditions as above), the assay was performed as described in Section 2 and showed the disappearance of styrene oxide compared with a blank without cells. Epoxide hydrolysis activity could even be observed in a mixture containing 33% or 50% acetone using whole cells of both microorganisms.

In principle the epoxide hydrolase assay based on NPB might interfere with other enzymatic activities such as glutathion-*S*-transferases or reductases present in whole microorganisms or crude lysed cells. However, this we exclude due to the correlation between results obtained by the NPB assay and GC analysis (see above).

The values obtained from the NBP assay were further verified by using whole or lysed cells from *E. coli* DH5α and boiled *Rhodococcus* sp. cells which showed no epoxide hydrolase depletion. Furthermore, no spontaneous reaction with nucleophilic compounds

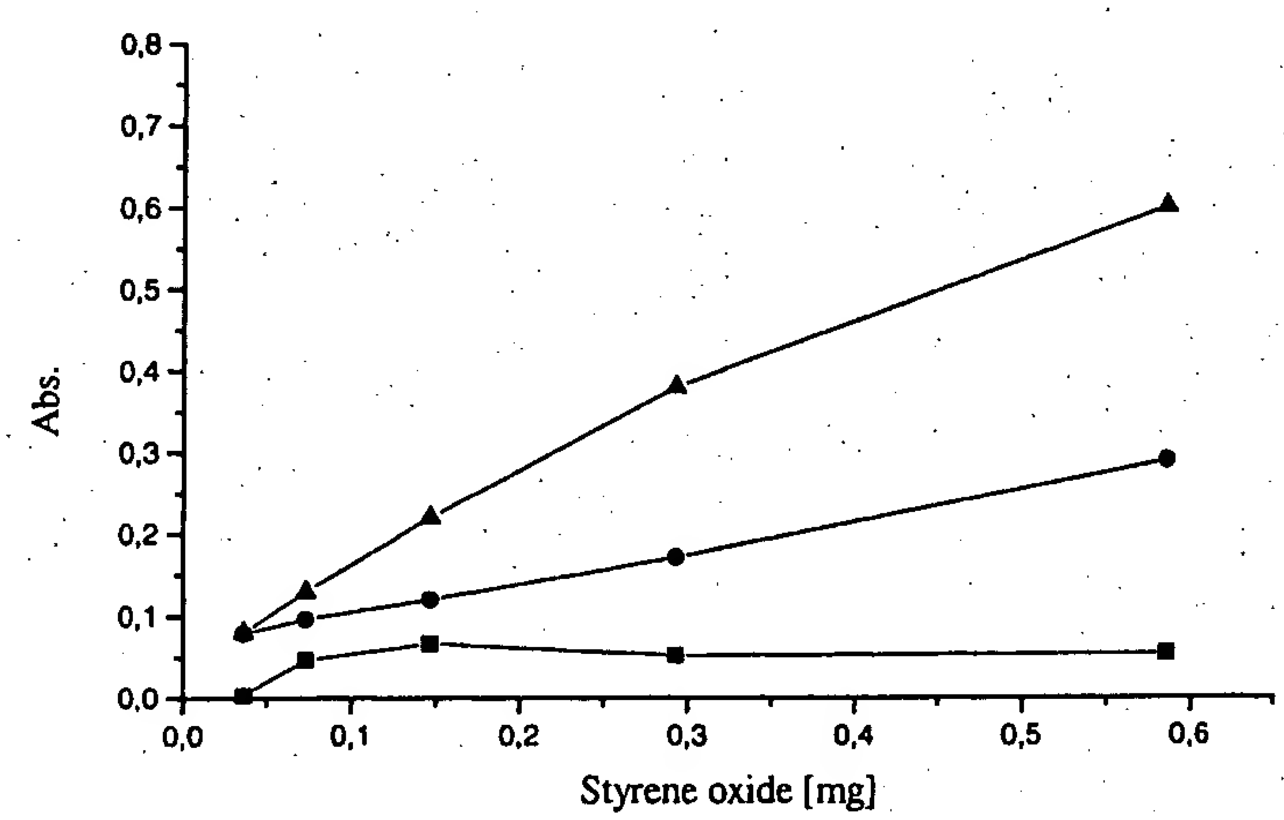


Fig. 4. Absorbance in the presence of *E. coli* DH5α after NBP assay (●), without cells/NBP (■) and after cell lysis/NBP (▲) at different concentrations of styrene oxide per well (45 min incubation, 39°C).

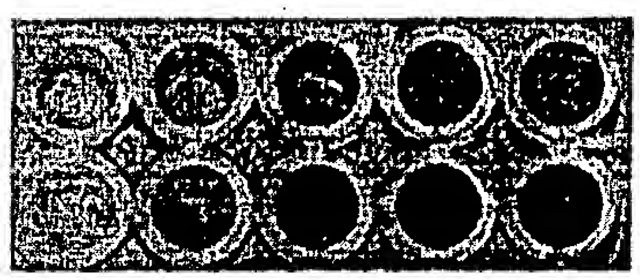


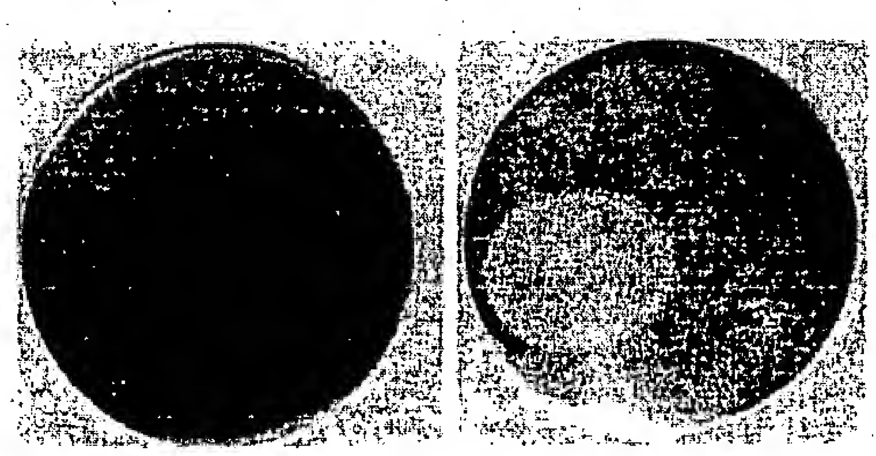
Fig. 5. Control experiment using whole (top row) and boiled (bottom row) cells from *Rhodococcus* sp. Increasing styrene oxide concentrations from left to right.

(e.g. water, proteins, DNA) in these mixtures under the assay conditions and within the reaction time could be observed (Fig. 5).

In order to screen for epoxide hydrolase activity directly on agar plates, a filter paper assay was used. Here, *Rhodococcus* sp. NCIMB 11216 or *E. coli* DH5α colonies of a gene bank containing genes from the same *Rhodococcus* sp. were grown on LB agar plates. The colonies were then transferred to filter paper preincubated with styrene oxide as given in Section 2. Epoxide hydrolase activity could easily be detected by the formation of colorless halos on the remaining blue filter paper caused by the hydrolysis of styrene oxide by the epoxide hydrolase from *Rhodococcus* sp. NCIMB 11216 (Fig. 6).

4. Conclusions

Two novel assay formats useful for screening of epoxide hydrolase activity in whole microorganisms (intact, lysed cells or containing gene banks encoding epoxide hydrolase activity) were developed. The microtiter-plate assay should also allow to rapidly and quantitatively monitor the activity of epoxide hydrolases during purification procedures. For both applications, a fast and accurate procedure is important, because large numbers of samples have to be analyzed in reasonable time. The microtiter-plate assay is easy to perform with the help of laboratory automation techniques and large libraries can be screened fast with high accuracy and reproducibility. The microtiter-plate assay can be applied to different epoxides such as styrene oxide, ethyl phenylglycidate, indene oxide or *n*-hexane oxide representing different classes of substituted epoxides. Both assays should also be transferable to other epoxides having sufficiently high alkylating strength. By the use of enantiomerically pure epoxides, it should also be possible to identify enzymes with high enantioselectivity and desired enantiopreference. Furthermore, the filter paper assay enables a direct screening on agar plates, which is especially suitable for the screening of large numbers of colonies without transferring them to microtiter-plates.



Control *Rhodococcus* sp.

Fig. 6. Filter paper assay without enzyme as control (left) and with styrene oxide after incubation in the presence of whole cells from *Rhodococcus* sp. (right).

References

[1] R.A. Johnson, K.B. Sharpless, in: I. Ojima (Ed.), Catalytic Asymmetric Synthesis, Verlag Chemie, New York, 1993, p. 103.
[2] T. Katsuki, K.B. Sharpless, J. Am. Chem. Soc. 102 (1980) 5974.
[3] S. Pedragosa-Moreau, A. Archelas, R. Furstoss, Bull. Soc. Chim. Fr. 132 (1995) 769.
[4] K. Faber, M. Mischitz, W. Kroutil, Acta Chem. Scand. 50 (1996) 249.
[5] F. Oesch, Xenobiotica 3 (1973) 305.
[6] M. Arand, W. Hinz, F. Müller, K. Hänel, L. Winkler, A. Mecky, H. Knehr, H. Dürk, H. Wagner, M. Ringhoffer, F. Oesch, Control. Mech. Carcinog. (1996) 116.
[7] B.D. Hammock, D. Grant, S. Storms, in: I. Sipes, C. McQueen, A. Gandolfi (Eds.), Comprehensive Toxicology, ch. 18, Epoxide hydrolases, Pergamon Press, Oxford, 1997, p. 283.
[8] M. Mischitz, K. Faber, A. Willets, Biotechnol. Lett. 17 (1995) 893.
[9] X.-J. Chen, A. Archelas, R. Furstoss, J. Org. Chem. 58 (1993) 5528.
[10] S. Pedragosa-Moreau, A. Archelas, R. Furstoss, J. Org. Chem. 58 (1993) 5533.
[11] W. Kroutil, M. Mischitz, P. Plachota, K. Faber, Tetrahedron Lett. 46 (1996) 8379.
[12] R. Rink, M. Fennema, M. Smids, U. Dehmel, D.B. Janssen, J. Biol. Chem. 272 (1997) 14650.
[13] E. Misawa, C.K. Chan Kwo Chion, I.V. Archer, M.P. Woodland, N.Y. Zhou, S.F. Charter, D.A. Widdowson, D.J. Leak, Eur. J. Biochem. 253 (1998) 173.

[14] R.B. Westkaemper, R.P. Hanzlik, Anal. Biochem. 102 (1980) 63.
[15] E.C. Dietze, E. Kuwano, B.D. Hammock, Anal. Biochem. 216 (1994) 176.
[16] J.C. Moore, F.H. Arnold, Nature Biotechnol. 14 (1996) 315.
[17] W.P.C. Stemmer, Nature 370 (1994) 389.
[18] R.B. Westkaemper, R.P. Hanzlik, Arch. Biochem. Biophys. 208 (1981) 195.
[19] R.N. Wixtrom, B.D. Hammock, Anal. Biochem. 174 (1988) 291.
[20] R. Preussmann, H. Schneider, F. Epple, Arzneimittel-Forsch. 19 (1969) 1059.
[21] K. Hemminki, T. Heinonen, H. Vainio, Arch. Toxicol. 49 (1981) 35.
[22] A. Barbin, J.-C. Béréziat, A. Croisy, I.K. O'Neill, Chem. Biol. Interact. 73 (1990) 261.
[23] E. Eder, T. Neudecker, D. Lutz, D. Henschler, Chem. Biol. Interact. 38 (1982) 303.
[24] J.H. Kim, J.J. Thomas, Bull. Environ. Contam. Toxicol. 49 (1992) 879.
[25] R. Serrentino, P.G. Gervati, Boll. Soc. Ital. Sper. 56 (1980) 2393.
[26] L.G. Hammock, B.D. Hammock, J.E. Casida, Bull. Environ. Contam. Toxicol. 12 (1974) 759.
[27] J. Sambrook, E.F. Fritsch, T. Maniatis, Molecular cloning: A laboratory manual, Cold Spring Harbor Laboratory, Cold Spring, New York, 1989.

YEAST
Yeast 15, 973–986 (1999)

Yeast Functional Analysis Reports

Chemotyping of Yeast Mutants Using Robotics

KLAUS-JÖRG RIEGER[1,2]*, MOHAMED EL-ALAMA[2]†, GEORG STEIN[3], CHARLES BRADSHAW[2]†, PIOTR P. SLONIMSKI[1] AND KINSEY MAUNDRELL[2]†

[1]*Centre de Génétique Moléculaire du Centre National de la Recherche Scientifique, Laboratoire Propre Associé à L'Université Pierre et Marie Curie, F-91198 Gif-sur-Yvette, France*
[2]*Geneva Biomedical Research Institute, Glaxo-Wellcome, 14 chemin des Aulx, CH-1228 Plan-les-Ouates, Geneva, Switzerland*
[3]*Institut für Botanik I, Abt. Botan. Cytologie und Morphologie, Heinrich-Heine-Universität Düsseldorf, Universitätsstr. 1, D-40225 Düsseldorf, Germany*

By now, the EUROFAN programme for the functional analysis of genes from the yeast genome has attained its cruising speed. Indeed, several hundreds of yeast mutants with no phenotype as tested by growth on standard media and no significant sequence similarity to proteins of known function are available through the efforts of various laboratories. Based on the methodology initiated during the pilot project on yeast chromosome III (*Yeast* **13**, 1547–1562, 1997) we adapted it to High Throughput Screening (HTS), using robotics. The first 100 different gene deletions from EUROSCARF, constructed in an FY1679 strain background, were run against a collection of about 300 inhibitors. Many of these inhibitors have not been reported until now to interfere *in vivo* with growth of *Saccharomyces cerevisiae*. In the present paper we provide a list of novel growth conditions and a compilation of 49 yeast deletants (from chromosomes II, IV, VII, X, XIV, XV) corresponding to 58% of the analysed genes, with at least one clear and stringent phenotype. The majority of these deletants are sensitive to one or two compounds (monotropic phenotype) while a distinct subclass of deletants displays a hyper-pleiotropic phenotype with sensitivities to a dozen or more compounds. Therefore, chemotyping of unknown genes with a large spectrum of drugs opens new vistas for a more in-depth functional analysis and a more precise definition of molecular targets.

KEY WORDS — drug-sensitivity/resistance; functional analysis; genome; robotics; *Saccharomyces cerevisiae*

INTRODUCTION

Currently, less than half of the genes of the *Saccharomyces cerevisiae* genome have been characterized, either genetically and/or biochemically. The majority of the remaining genes have no significant similarity to genes of known function. In addition, deletion of these genes often does not lead to obvious growth defects or other detectable phenotypic characteristics, as tested by growth on standard media. The search for phenotypes by screening yeast mutants against a standardized compound collection is a logical step to increase our knowledge of a particular gene and its function.

Recently we reported a microtiterplate-based screening for phenotypes of functionally uncharacterized genes from yeast chromosome III (Rieger *et al.*, 1997, 1998). In the framework of the EUROFAN programme, we have now extended our previous approach in two ways: first, we have used

*Correspondence to: K.-J. Rieger, Centre de Génétique Moléculaire du Centre National de la Recherche Scientifique, Avenue de la Terrasse, F-91198 Gif-sur-Yvette cedex, France. Tel.: 33-1-69 82 31 78; fax: 33-1-69 07 55 39.
†Present address: Serono Pharmaceutical Research Institute, 14 chemin des Aulx, CH-1228 Plan-les-Ouates, Geneva, Switzerland.
Contract/grant sponsor: EUROFAN Programme of the EC; Contract/grant number: BIO4-CT95-0080.
Contract/grant sponsor: EC; Contract/grant number: BIO4-CT95-0080.

CCC 0749–503X/99/10B973–14 $17.50

Received 6 October 1998
Accepted 17 December 1998

an enlarged panel of inhibitors and growth conditions; and second, we have adapted the assay for high throughput by using a robotic workstation. This approach has been used in the framework of the consortium B1 to analyse the first 100 deletants furnished by EUROSCARF (European Saccharomyces Cerevisiae Archives for Functional Analysis; deletants 10 001–10 096). In addition to various stress conditions and inhibitors listed in the preceding publication, we have analysed the inhibitory activity of more than 300 new chemicals on yeast, which in many cases have never been tested on *S. cerevisiae* (Rieger *et al.*, 1997, 1998).

The method used gives a large number of phenotypes, which will be referred to henceforth as 'chemotypes'. We have found that over 50% of the deletants listed in the MIPS/EUROSCARF database as having no detectable phenotype, do have significantly altered growth characteristics when exposed to one or more of the compounds/conditions in our standard set.

MATERIALS AND METHODS

Yeast strains

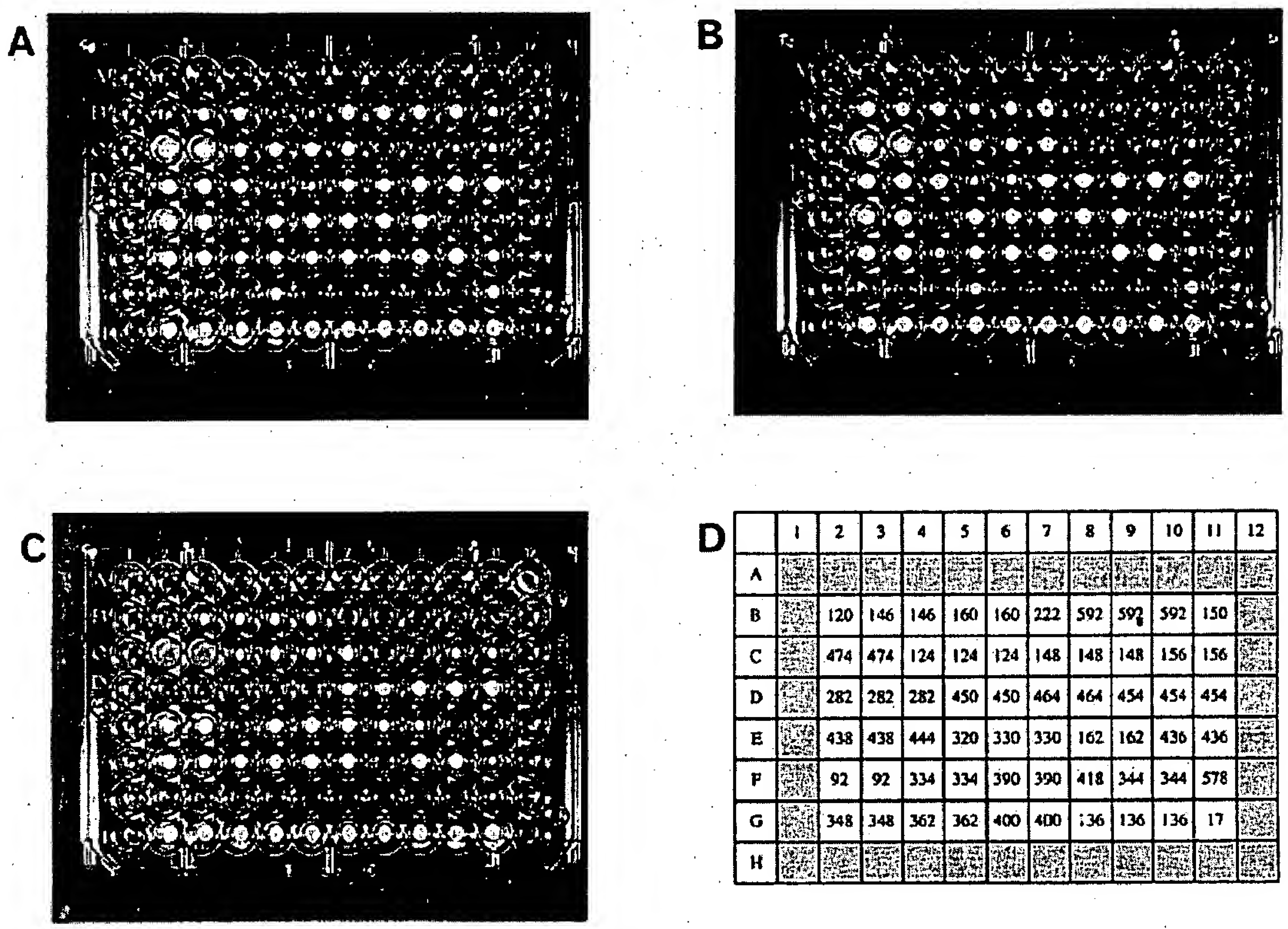
All mutant strains generated during the EUROFAN project were provided by EUROSCARF stock center (Frankfurt, Germany; for address and persons to contact, see Acknowledgements). All targeted gene deletions were carried out in the strain FY1679 (deletants FY 10001A/B-10096 A/B). For the primary phenotype screening, only mutants with the alpha mating type were used (FY 10001 B-10096 B). As reference strain we utilized FY1679-18B (Baudin *et al.*, 1993; Rieger *et al.*, 1997). 'Hits' (sensitivity or resistance to a given compound) were screened a second time against an independent set of the same mutants to eliminate false positives. As a further check, we tested random 'hits' discovered by our approach by showing co-segregation of the chemotype with the deletion marker (tetrad analysis) and complementation by the corresponding wild-type gene.

Analysed strains were as follows: YBL047c/YBR008c, YBR014c, YBR016w, YBR041w, YBR042c, YBR043c, YBR161w, YBR162c, YBR175w, YBR182c, YBR180w, YBR260c, YBR264c//YDL074, YDL117w, YDL223c, YDL237w, YDL238c, YDL239c//YGL186c/YGR136w, YGR221c, YGR223c, YGR224w, YGR225w, YGR226c, YGR231c, YGR232w//YJL038c, YJL045w, YJL046w, YJL047c, YJL048c, YJL049w, YJL051w, YJL055w, YJL056c, YJL057c, YJL058c, YJL059w, YJL062w, YJL065c, YJL169w, YJL199c, YJL201w, YJL204c, YJL206c, YJL207c, YJL213w//YLR036c, YLR042c//YNL054w, YNL056w, YNL058c, YNL063w, YNL065w, YNL066w, YNL196c, YNL200c, YNL206c, YNL208w, YNL211c, YNL212w, YNL213c, YNL214w, YNL215w, YNL217w, YNL218w, YNL281w, YNL283c, YNL285w, YNL286w, YNL288w//YOL018c, YOL087c, YOL088c, YOL091w, YOL098c, YOL101c, YOL104c, YOL105c, YOL132w/YOR311c, YOR322c.

Chemicals

Stock solutions of the different compounds were made mainly in dimethylsulphoxide (DMSO), NaOH and water (some in ethanol, methanol). Stock solutions were filter-sterilized and stored following the instructions of the suppliers (Alexis, Calbiochem, Research Biochemicals International, Sigma, Tocris Cookson). All of the chemicals were of the highest available purity grade. Final concentrations are given in brackets:

002, ethylenediamine-tetraacetic acid (EDTA) [0·7/0·6 mM]; **004**, tungstic acid [35/27·5 mM]; **006**, $CdCl_2$ [55/50 μM]; **007**, CsCl [100/90 mM]; **008**, $CoCl_2$ [0·6/0·5 mM]; **009**, $CuSO_4$ [9/8·75/6·75 mM]; **010**, $NiCl_2$ [1·25/1·15 mM]; **017**, $MgCl_2$ [490 mM]; **020**, RbCl [500/300 mM]; **021**, $SrCl_2$ [250 mM]; **022**, LiCl [160/150 mM]; **023**, $MnCl_2$ [3·5/3 mM]; **024**, $ZnCl_2$ [3·5/3 mM]; **032**, nalidixic acid [0·6/0·5 mg/ml]; **040**, sodium-o-vanadate [2·6/2·3 mM in 50 mM KOH; tested with 704 as growth medium]; **044**, 2,2 dipyridyl [10/9/7·5 μg/ml]; **050**, cycloheximide [0·18/0·16/0·14 μg/ml]; **054**, benomyl [25/20 μg/ml; tested on Petri dishes]; **056**, 1,10 phenanthroline [15/12·5 μg/ml]; **070**, fluorescent brightener 28 [2/1·5/1 mg/ml; tested on Petri dishes]; **073**, *p*-chloromercuribenzoic acid (PCMB) [0·24/0·22/0·21 mM]; **078**, diltiazem–HCl [70/65/60 μg/ml]; **092**, benzamidine [33/31 mM]; **098**, chlorpromazine [60/50/40/30 μM in DMSO]; **100**, quercetin [5/3 mM, in ethanol]; **146**, valeryl salicylate [2·75/2·5 mM in DMSO]; **148**, loperamide [1·4/1·2/1 mM in DMSO]; **150**, 6-dimethylaminopurine [7 mM in DMSO]; **156**, 2,4 dinitrophenol [1·3/1·1 mM in DMSO]; **160**, sanguinarine [0·15/0·13 mM in DMSO]; **170**, trimethoprim [3·5 mM in DMSO]; **180**, tetrindole mesylate [9/8·5/8/7 μM in DMSO]; **222**, nifedipine [2·25 mM in DMSO]; **224**, nordihydroguaiaretic



	1	2	3	4	5	6	7	8	9	10	11	12
A												
B		120	146	146	160	160	222	592	592	592	150	
C		474	474	124	124	124	148	148	148	156	156	
D		282	282	282	450	450	464	464	454	454	454	
E		438	438	444	320	330	330	162	162	436	436	
F		92	92	334	334	390	390	418	344	344	578	
G		348	348	362	362	400	400	136	136	136	17	
H												

Figure 1. Chemotyping of *S. cerevisiae* deletants in microtitre plates. Shown are representative examples of mutants provided by the EUROFAN project. Preparation of medium, inhibitors, cell suspensions and plate filling were done as outlined in Materials and Methods. Each microtitre plate was seeded with one deletant and analysed in 60 different growth conditions (corresponding to 31 inhibitors at one or several different concentrations in the examples shown above) per plate. The outer rows of wells were filled up with agar but were not used for chemotyping. The bottom line in each plate (2H–11H) corresponds to 10 replicas of the deletant in the absence of inhibitor ('internal control'). Microtitre plates with control or deleted strains (all deletions were made in a FY1679 background and all strains were *MATα*) were incubated for various periods of time at 28°C. Mutant strains were as follows: (A) YNL215w; (B) YOR322c; (C) YJL204c. Panel (D) refers to the position of the corresponding inhibitors in each microtitre plate (compounds that are not listed in Materials and Methods: zaprinast [120], *t*-butylhydroquinone [124], disulphiram [136], mycophenolic acid [162], N_1, N_{12}-diethylspermine tetrahydrochloride [320], 2,5 anhydro-D-mannitol [330], N-ethylmaleimide [344], benserazide [400], 1-*O*-hexyl-rac-glycerol [464]). Photographs were taken after 2 days of incubation at 28°C.

acid (NDGA) [0·8 mM in DMSO]; **246**, canthardate [0·5 mM in DMSO]; **276**, ruthenium red [2·5 mM in DMSO]; **282**, phenyl-ethylamine [0·40/0·035/0·03% in DMSO]; **310**, thiabendazole [100/80 μg/ml in DMSO; tested on Petri dishes]; **334**, doxorubicin [0·1/0·08 mM]; **348**, diethyl maleate [0·15/0·14/0·13% in ethanol]; **356**, methyl caffeate [2·75/2·5 mM in DMSO]; **362**, polymyxin B [0·11/0·1 mM]; **378**, latrunculin B [0·03/0·0275 mM in DMSO]; **390**, 4-aminopyridine [6/5 mM in DMSO]; **418**, β-chloro-L-alanine [9 mM in DMSO]; **436**, azaserine [0·28/0·26/0·24 mM]; **438**, BAPTA-tetrapotassium salt [1·25/1·0 mM]; **444**, guanosine 5′-O-(2-thiodiphosphate) [0·7 mM in DMSO]; **450**, diphenyleneiodonium [0·15/0·13 mM in DMSO]; **454**, hexadecylphosphocholine [2/1·5/1 μM in DMSO]; **474**, compound 48/80 [0·7/0·6 mM]; **476**, hydroxyurea [10 mg/ml]; **484**, caffeine [14/13/10 mM]; **486**, tunicamycine [0·14/0·12% in DMSO]; **578**, daunorubicin [0·1265 mM in DMSO]; **592**, cerulenin [2·25/2/1·75 μM in DMSO]; **598**, isopropyl (*N*-3-chloro-phenyl)-carbamate (IPCPC) [0·4/0·2 mM in DMSO]; **701**, tetraethylammonium chloride [55/40/20 μg/ml].

Table 1. Selective list of chemicals and growth conditions for the phenotypic analysis of genes of unknown function from EUROSCARF (for strains, see Materials and Methods). Numbers refer to the preparation of the corresponding media as outlined in Materials and Methods. Literature quotations are non-exhaustive—for simplicity we often refer to the information provided by the supplier (Alexis, 1997)—and indicated only for substances that have not been described earlier (Rieger *et al.*, 1997).

Compound	Function and/or target and mode of function of inhibitors
004 Tungstic acid	Potent inhibitor of xanthine oxidase, inhibits sulfite oxidase in rats (Johnson *et al.*, 1974); competitive antagonist of molybdenum in animals; phosphotyrosine phosphatase (PTP) inhibitor (mouse); inhibitor of acid phosphatase (human)
092 Benzamidine	Peptidase inhibitor
098 Chlorpromazine	Inhibitor of calmodulin stimulation of cyclic nucleotide phosphodiesterase; dopamine antagonist antiemetic and antipsychotic; also acts as peripheral vasodilator; inhibits TNF-α production; potent PLA_2 inhibitor, inhibits nitric oxide synthase in mouse brain and prevents lipopolysaccharide induction of nitric oxide synthase in murine lung (Alexis, 1997); counteracts alpha factor-induced growth arrest in G_1 of the *S. cerevisiae* cell cycle (Ruiz and Rodriguez, 1986); acts as photomutagen (Chetelat *et al.*, 1993)
100 Quercetin	Antioxidant flavonoid; inhibitor of mitochondrial ATPase, cAMP- and cGMP phosphodiesterases; inhibitor of protein tyrosine kinases and protein kinase C; induces apoptosis in K562, Molt-4, Raji and MCAS tumour cell lines (Alexis, 1997)
146 Valeryl salicylate	Selective inhibitor of COX-1 (Alexis, 1997)
148 Loperamide	Ca^{2+}-channel antagonist; binds to opioid receptors; has anti-diarrhoeal activity at low concentrations; binds to calmodulin at high concentrations; meperidine analogue (Alexis, 1997)
150 6-Dimethylaminopurine	Non-selective inhibitor of cdc2 and other protein kinases (Alexis, 1997; Chong *et al.*, 1995; Felix *et al.*, 1989)
160 Sanguinarine	Na^+/K^+- and Mg^{2+}-ATPase inhibitor (Alexis, 1997)
170 Trimethoprim	Inhibitor of dihydrofolate reductase (Bertani and Campbell, 1994)
180 Tetrindole mesylate	Novel antidepressant; selective monoamine oxidase A inhibitor
222 Nifedipine	Calcium channel blocker (L-type); vasodilator (Alexis, 1997)
224 Nordihydroguaiaretic acid	Antioxidant; lipoxygenase inhibitor (Alexis, 1997; Jensen *et al.*, 1992; Kustimur *et al.*, 1997)
246 Canthardate	High potency inhibitor of protein phosphatase 2A (Alexis, 1997)
276 Ruthenium red	Capsaicin and calcium antagonist; inhibitor of Ca^{2+}/Mg^{2+} ATPase (Alexis, 1997); channel blocker (Calvert and Sanders, 1995); affects wall morphogenesis in yeasts (Poli *et al.*, 1995)
282 Phenylethylamine	Potentiates dopamine and noradrenaline function in the central nervous system (Yu, 1994)
310 Thiabendazole	Microtubule depolymerizing drug (Castano *et al.*, 1996)
334 Doxorubicin	Antitumour antibiotic; inhibitor of reverse transcriptase and RNA polymerase; immunosuppressive agent; highly effective myotoxin that inhibits topoisomerase II; binds to nucleic acids, presumably by specific intercalation into the DNA double helix, thereby inhibiting nucleic acid synthesis; induces apoptosis (Alexis, 1997; Kule *et al.*, 1994; Patel *et al.*, 1997)
348 Diethyl maleate	Generates oxidative stress (Kuge *et al.*, 1997)
356 Methyl caffeate	Inhibitor of ornithine decarboxylase and protein tyrosine kinase (Alexis, 1997)
362 Polymyxin B	Inhibitor of protein kinase C; antibiotic; breaks bacterial membranes by incorporating into the phospholipid of the outer membrane and activating phospholipase (Alexis, 1997; Boguslawski, 1992)

Table 1. *Continued.*

Compound	Function and/or target and mode of function of inhibitors
378 Latrunculin B	Structurally unique marine toxin; actin filament modulator; 10–100-fold more potent than cytochalasins; whereas cytochalasin D induces dissolution of F-actin and stress fibre contraction in fibroblasts in culture, latrunculin B causes a shortening and thickening of stress fibres. These differences may indicate that the two classes of compounds have distinct target sites (Alexis, 1997)
390 4-Aminopyridine	Induces depolarization of GABA neurons; prolongs action potential in demyelinated nerve fibres (Alexis, 1997)
418 β-Chloro-L-alanine	Inhibitor of alanine aminotransferase (Morino *et al.*, 1979)
436 Azaserine	Potent inhibitor of purine biosynthesis; glutamine antagonist (Aronow *et al.*, 1986; Becker and Kim, 1987)
438 BAPTA	Highly selective calcium chelating agent (Alexis, 1997; Loukin and Kung, 1995)
444 Guanosine 5′-*O*-(2-thiodiphosphate)	Non-hydolyzable GDP-analogue that competitively inhibits G-protein activation by GTP and GTP analogues (Alexis, 1997)
450 Diphenyleneiodonium	Binds strongly to flavoproteins and is thus a powerful and specific inhibitor of several important enzymes, including nitric oxide synthase, NADH-ubiquinone oxidoreductase, NADPH oxidases and NADPH cytochrome P450 oxidoreductase; recently, nitric oxide synthase, which shows significant homology with cytochrome P450 reductase, has shown to be irreversibly inhibited by this compound (Alexis, 1997; Lesuisse *et al.*, 1996)
454 Hexadecylphosphocholine	Potent antitumour agent; inhibits protein kinase C and phosphatidylcholine biosynthesis; co-stimulates human T cell activation (Alexis, 1997)
474 Compound 48/80	Blocks calmodulin and human platelet phospholipase C; blocks ADP ribosylation; activates G-proteins (likely G_1 proteins) in a manner similar to G protein-coupled receptors; modulates human platelet phospholipase A_2 in a concentration-dependent, biphasic manner; stimulates release of histamine and arachidonic acid (Alexis, 1997)
578 Daunorubicin	Potent anticancer agent whose potential target site may be mitochondrial cytochrome C oxidase; inhibits RNA and DNA synthesis; also inhibits eukaryotic topoisomerase I and II; induces DNA single-strand breaks and apoptosis in HeLaS3 tumor cells (Alexis, 1997; Kule *et al.*, 1994)
598 Isopropyl (*N*-3-chlorophenyl)-carbamate	Mitotic poison, inhibits plant metabolism

Non-standard culture conditions

702, slow-growth (three growth temperatures: 16, 28, 36°C; tested on YPD); **703**, pH 3·41/3·8/4·16 (concentrated YPD [90% of final volume] was mixed at about 60°C with filter-sterilized × 10 acetate buffer (1 M) of pH as indicated); **704**, N3 (standard glycerol medium: 1% yeast extract, 1% bactopeptone, 2% glycerol, 0·05 M sodium phosphate (pH 6·2, 100 ml/l) and 80 mg/l adenine); **705**, pH 7·8/8 (concentrated YPD [90% of final volume] was mixed at about 60°C with filter-sterilized × 10 phosphate buffer [K_2HPO_4/KH_2PO_4, 1 M] of pH as indicated).

Establishing the range of inhibitor concentrations of the reference strain

To establish the threshold concentration (or a range of concentrations) for the reference strain, compounds were serially (× 4) diluted in 20% DMSO in microtitre plates. Where necessary, the final inhibitor concentration was determined more precisely by testing concentrations between the

last dilution (e.g. × 4) that completely inhibited growth and the next serial dilution (e.g. × 16) allowing growth. In all cases plates were filled with compounds (20 µl/well) followed by distribution of YPD medium (180 µl/well, 0·7% agar). Plates were stored for at least 24 h at 4°C to allow chemicals to diffuse into the agar prior to dispension of the control-cells (FY1679-18B). The reference strain was cultured overnight in liquid YPD at 28°C, dilutions were made in YPD and about 1000 cells/well (20 µl) were spotted in the wells. Cell growth was then followed for up to 7 days in 28°C in the presence or absence of the compounds.

Media, replication of compound arrays by a robotic workstation and cell culture conditions

The majority of assays was done with standard complete glucose medium (YPD, 1% yeast extract [Difco Laboratories, Detroit, USA], 1% bactopeptone [Difco], 2% glucose, 20 mg/l adenine; growth medium was solidified by adding 0·7% Bacto-Agar [Difco]). In some cases chemotypes were scored on Petri dishes (e.g. growth conditions 54, 310, 703, 705). In this case, medium was solidified by adding 2% Bacto-Agar.

Compounds arrays were replicated at 20 µl per well into sterile 96-well microtitre plates (NUNC, Intermed, Polylabo, Switzerland) from larger volumes held in 3 ml deep-(96)-well 'master' plates (Quiagen). This operation was carried out using a Genesis RSP100 pipetting robot (TECAN AG, Switzerland). Sterile 96-well microtitreplates and 'master' plates were stored with loose lids on a high capacity storage carousel. Plates were transferred between the carousel and the pipettor using a CRS A255 robotic arm (CRS, Canada). Movements were programmed using CLARA dynamic scheduling software (SCITEC, Lausanne, Switzerland). Each of these 'master' plates contained sufficient volumes of compound solutions to produce up to 54 replicated plates. The peripheral wells were left empty, to avoid previously observed 'edge' effects. Hot growth medium (YPD, containing 0·7% agar) was subsequently dispensed at 180 µl per well, into all 96 wells of each of the replicated compounds plates using a Multidrop-831 eight-channel dispenser (Labsystems OY, Finland). This was carried out as a manual batch operation, using a custom-modified delivery tube assembly to deliver agar at 50–65°C approximately. Tubing was sterilized with ethanol and rinsed with sterile deionized water immediately before priming with the sterile agar solution. Plates were then 'loose-lidded' and allowed to cool at 4°C for at least 1 day in a horizontal position before manually dispensing yeast cell suspensions using a handheld semiautomatic multichannel pipette EDP-Plus M8 (Rainin Instrument Company, Inc., Woburn, MA).

Strains were cultured overnight in liquid YPD at 28°C. On the next day, the optical density of all cultures was determined and cell density adjusted to 5 × 10 000 cells/ml by dilution into fresh YPD. For the seeding of cells in microtitreplates, cell suspensions were vortexed vigorously and filled at about 1 ml in the tubes of a cluster tube 8-strip rack (Costar, Polylabo, Paris). 20 µl of the cell suspension, corresponding to about 1000 cells/well, were inoculated into the wells and the microtitre plate placed on a shaker for 5 s in order to cover the agar surface uniformly. Plates were then incubated at 28°C for up to 10 days. From the first day of incubation on, growth of the mutant strains was scored visually, either directly on the plate or later on photographs, by comparison with the growth of the corresponding control strains.

RESULTS AND DISCUSSION

In the present paper we describe the systematic phenotypic analysis of 85 yeast mutants using a robotic workstation in combination with a battery of some 300 growth inhibitory conditions. Mutants as a starting material for functional analysis were generated during the EUROFAN 1 programme and have been obtained from EUROSCARF. These mutants, which had no obvious phenotypes (as tested by growth on standard media: YPD, respiratory medium, different temperatures), were therefore analysed by screening them against a collection of chemical compounds. This library contained beside the growth conditions described earlier (Rieger *et al.*, 1997) a large number of 'novel' inhibitors, that have not been reported previously in the literature to interfere with cell growth of *Saccharomyces cerevisiae*. Some representative examples of the screening are given in Figure 1. Among the compounds, only those that gave a clear and stringent phenotype with the analysed set of mutants are presented here (Table 1). Many novel inhibitors (e.g. isopropyl *N*-(3-chlorophenyl)-carbamate, loperamide, sanguinarine, chlorpromazine, ruthenium red,

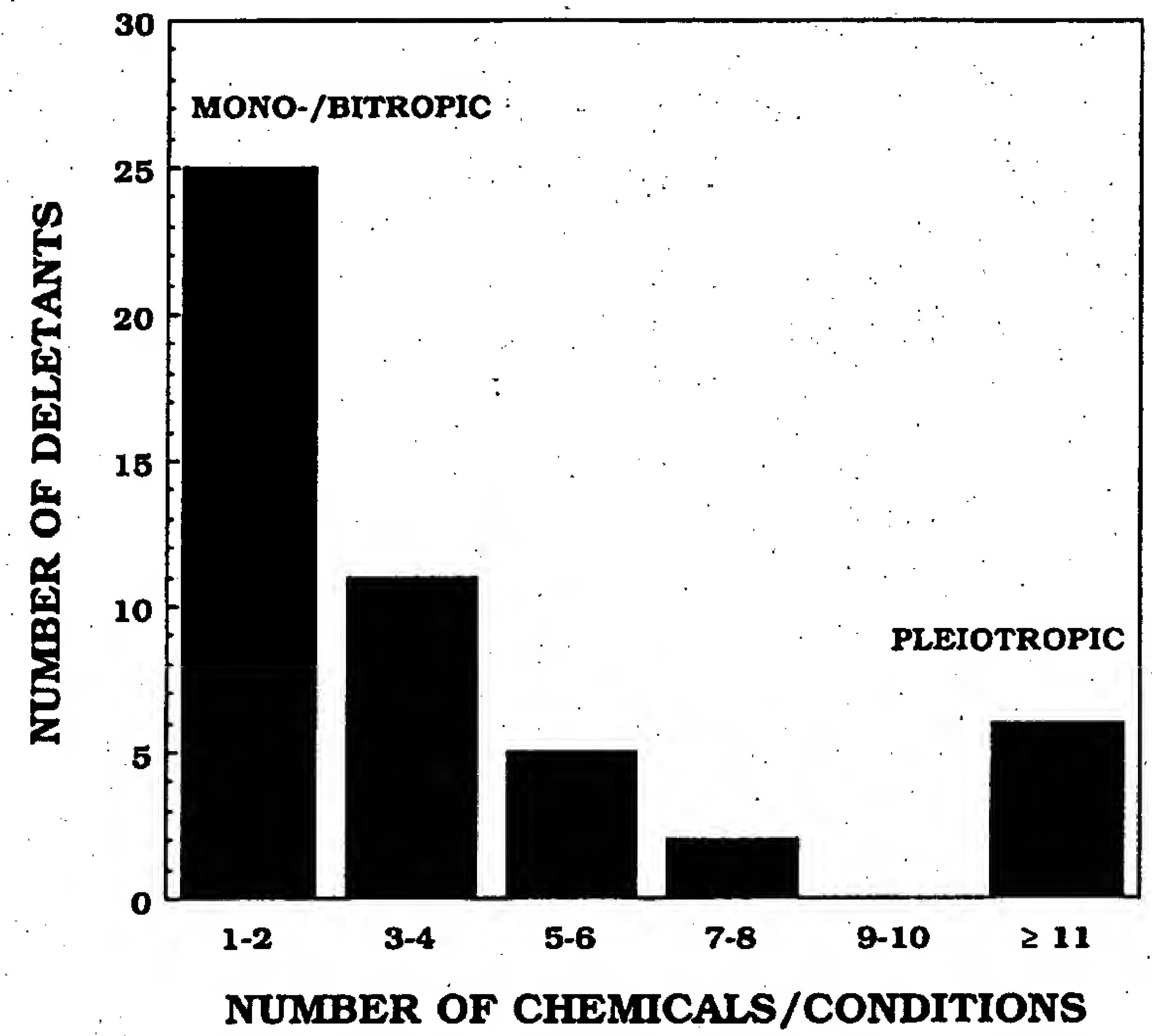


Figure 2. Distribution of the number of deletants as a function of the number of chemotypes. For each deletant the number of chemotypes was determined (see Table 2): e.g. YGR224w had one chemotype (hypersensitivity to polymyxin B) while YOL088c had three chemotypes (hypersensitivity to $CdCl_2$ and sensitivity to cerulenin and loperamide). The deletants were grouped into classes displaying 1 or 2 chemotypes, 3 or 4 chemotypes, etc. and the number of deletants in each class is shown.

azaserine, guanosine 5′-*O*-2-(thiodiphosphate), etc.) turned out to be highly discriminating, since in many instances specific deletants were much more sensitive to the drug than the reference strain.

Interestingly, as shown in Figure 2, the deletants can be grouped in two classes. In the first class, the majority of the mutants displayed sensitivity to a small number of inhibitors/chemicals: most of them were sensitive to one or two inhibitors tested (e.g. YOR311c, YGR224w, YBR043c; see Table 2) while progessively fewer mutants were sensitive to a larger number of compounds tested. Obviously (see Table 2), the chemicals to which those mono-/bitropic deletants are sensitive are characteristic and specific for each mutant (e.g. YBR043c is hypersensitive to $MnCl_2$, while YNL058c is sensitive to loperamide). It is possible that this class comprises genes which are involved in rather specific cellular processes. In addition their deletion does not severely interfere with the 'general health' of yeast. On the other hand, the second class, which is clearly distinct from the first, displays highly pleiotropic phenotypes, being sensitive to at least 11 different inhibitors/chemicals (in one extreme case, the mutant [YJL204c] was sensitive to more than 20 inhibitors/chemicals). It is plausible that the latter class comprises genes involved in some major cellular processes which hierarchically control several interconnected pathways like signal transduction, regulation of transcription, stress response or elaboration of cell architecture. Such multiple effects are understandable, since the primary defect can lead by a cascade of successive interactions to a variety of end effects which are indicative of the importance of the deleted gene. The panoply of end effects can be considered as 'symptoms' of the mutation and they can be grouped into 'syndromes' diagnostic of the biochemical/physiological process which is affected. This kind of analysis is beyond the scope of this article, which is essentially methodological, and focuses simply on the discovery of novel and

Table 2. List of chemotypes of gene deletions provided by EUROSCARF (Nos 10 001–10 096).

	ORF/gene name: comments (MIPS/SGD/YPD)	Phenotypes (this study): hypersensitivity [HS]; sensitivity [S]; resistance [R]
YOL088c/MPD2	Protein of unknown function; weak similarity to disulphide isomerases and ER60 proteases; has ER-targeting sequence (Zumstein *et al.*, 1995); overproduction suppresses lethality of protein disulphide isomerase depletion (Tachikawa *et al.*, 1997)	Cadmium chloride [HS] (6), cerulenin [S] (592), Loperamide [S] (148)
YOL087c	Protein of unknown function; has one WD (WD-40) domain; similarity to *S. pombe* hypothetical protein	1,10 phenanthroline [HS] (56), sanguinarine [HS] (160), NDGA [HS] (224), IPCPC [S] (598), gunaosine 5'-*O*-(2-thiodiphosphate) [S] (444)
YOR322c	Protein of unknown function; null mutant shows osmotic sensitivity on YEPD at 37°C (Pearson *et al.*, 1998); similarity to hypothetical *S. pombe* protein SPAC1F2.05	EDTA [S] (2), cadmium chloride [HS] (6), cycloheximide [HS] (50), 6-dimethylaminopurine [HS] (150), methyl caffeate [HS] (356), β-chloro-L-alanine [S] (418), caffeine [HS] (484), cerulenin [S] (592)
YOR311c	Protein of unknown function; similarity to *S. pombe* hypothetical protein	Nickel chloride [R] (10), nalidixic acid [R] (32)
YGR224w	Strong similarity to drug resistance protein *SGE1*, belongs to cluster II of the MFS-MDR family (Goffeau *et al.*, 1997)	Polymyxin B [HS] (362)
YBR043c	Similarity to benomyl/methotrexate resistance protein; belongs to cluster I of the MFS-MDR family (Goffeau *et al.*, 1997)	Manganese chloride [HS] (23)
YNL066w/SUN4	Involved in the ageing process [Mutants have longer lifespan and better viability upon starvation (Austriaco, 1996)]; strong similarity to *YIL123w*, *Uth1p*, *Nca3p* and *C. wickerhamii* β-glucosidase protein	Polymyxin B [S] (362), azaserine [S] (436)
YNL063w	Protein of unknown function; weak similarity to *Mycoplasma* protoporphyrinogen oxidase	Canthardate [S] (246), polymyxin B [S] (362), tetraethylammonium chloride [S] (701)
YNL058c	Protein of unknown function; similarity to *YIL117c*	Loperamide [S] (148)
YNL056w	Protein of unknown function; similarity to *YNL099c* and *SIW14* (protein tyrosine phosphatase; null mutant fails to show cell cycle arrest upon nutritient starvation, is sensitive to 5 mM caffeine and 1 M NaCl and shows delocalized actin upon nutrient starvation	Diltiazem–HCl [S] (78), caffeine [HS] (484)
YNL054w/VAC7	Integral vacuolar membrane protein, involved in vacuole morphology and inheritance (Bonangelino *et al.*, 1997)	EDTA [S] (2), copper sulphate [S] (9), cycloheximide [HS] (50), benomyl [HS] (54), fluorescent brightener [S] (70), azaserine [HS] (436), caffeine [HS] (484), slow-growth (702)

Table 2. *Continued.*

	ORF/gene name: comments (MIPS/SGD/YPD)	Phenotypes (this study): hypersensitivity [HS]; sensitivity [S]; resistance [R]
YBR260c	Protein of unknown function; similarity to *C. elegans* GTPase-activating protein, null mutant has slightly decreased viability during stationary phase; interacts genetically with *SLG1* (plasma membrane protein required for maintenance of cell wall integrity and for the stress response)	Cadmium chloride [HS] (6), diltiazem–HCl [S] (78), tunicamycine [HS] (486)
YNL215w	Protein of unknown function; similarity to hypothetical *S. pombe* protein	Cycloheximide [HS] (50), quercetine [S] (100), azaserine [S] (436), hydroxyurea [S] (476), IPCPC [S] (598), slow-growth (702)
YNL214w/PEX17	Component of the peroxisomal protein translocation-peroxin; null mutants lack morphologically detectable peroxisomes (Huhse *et al.*, 1998)	doxorubicin [S] (334), compound 48/80 [S] (474) daunorubicin [S] (578)
YBR264c/YPT10	Protein of unknown function; protein with similarity to rab proteins and other small GTP-binding proteins (Du *et al.*, 1998)	Chlorpromazine [S] (98), tetraethylammonium chloride [S] (701), pH 3·41/4·16 [S] (703)
YNL213c	Protein of unknown function; lysine-rich	Extreme slow growth according to EUROSCARF [HS] (702), pH 3·8 [S] (703)
YJL207c	Protein of unknown function; weak similarity to rat omega-conotoxin-sensitive calcium channel alpha-1 subunit rbB-I; weak similarity to *Spc98p*	Diltiazem–HCl [S] (78), chlorpromazine [HS] (98)
YJL204c	Protein of unknown function; weak similarity to *TOR2p* bud lacks consensus sequence for lipid kinases; member of a family of glycosyl hydrolases	Cesium chloride [HS] (7), cobalt chloride [HS] (8), rubidium chloride [S] (20), zinc chloride [HS] (24), cycloheximide [HS] (50), PCMB [HS] (73), diltiazem–HCl [HS] (78), chlorpromazine [HS] (98), quercetin [S] (100), loperamide [HS] (148), trimethoprim [HS] (170), tetrindole mesylate [S] (180), nifedipine [S] (222), NDGA [S] (224), ruthenium red [HS] (276), latrunculin B [S] (378), polymyxin B [HS] (362), β-chloro-L-alanine [S] (418), guanosine 5-*O*-(2-thiodiphosphate) [S] (444), diphenyleneiodonium [S] (450), slow-growth [S] (702), pH 4·16-sensitive [S] (703), pH 7·8/8-sensitive [HS] (705)
YOL018c/TLG2	Member of the syntaxin family of t-SNAREs; mutants show endocytosis defect and loss of *Kex2p*; affects late Golgi compartment (Holthuis *et al.*, 1998)	EDTA [HS] (2), cesium chloride [HS] (7), rubidium chloride [HS] (20), strontium chloride [HS] (21), lithium chloride [S] (22), cycloheximide [HS] (50), diltiazem–HCl [HS] (78), chlorpromazine [HS] (98), loperamide [HS] (148), NDGA [HS] (224), ruthenium red [HS] (276), latrunculin B [S] (378), BAPTA [S] (438), caffeine [HS] (484)
YNL196c/SLZ1	Sporulation specific protein, regulated by the transcription factor *Ume6* and expressed early in meiosis	Loperamide [HS] (148)

Table 2. *Continued.*

	ORF/gene name: comments (MIPS/SGD/YPD)	Phenotypes (this study): hypersensitivity [HS]; sensitivity [S]; resistance [R]
YOL105c/WSC3	Cell wall integrity and stress response component 3; *slg1* (*wsc1*)-null *wsc3*-null double mutant shows a lysis defect on YPD at room temperature and heat shock sensitivity (Verna *et al.*, 1997)	Canthardate [S] (246)
YNL200c	Protein of unknown function; strong similarity to human TGR-CL10C (thyroidal receptor for *N*-acetylglucosamine); contains a possible secretory signal; belongs to a group of 16 genes that are coordinately induced early in the diauxic shift (De Risi *et al.*, 1997)	Polymyxin B [HS] (362), azaserine [S] (436)
YNL206c	Protein of unknown function; protein with similarity to SSRP proteins (DNA structure-specific recognition protein); contains one PEST-site	Cesium chloride [S] (7), zinc chloride [S] (24), sodium-*O*-vanadate [R] (40), cycloheximide [HS] (50), benzamidine [S] (92), loperamide [HS] (148), sanguinarine [HS] (160), ruthenium red [S] (276), thiabendazole [HS] (310), azaserine [S] (436), caffeine [HS] (484), IPCPC [HS] (598), slow-growth [S] (702), pH 4·16 [S] (703)
YNL217w	Protein of unknown function; weak similarity to *E. coli* bis (5'-nucleosyl)-tetraphosphatase; contains a prokaryotic membrane lipoprotein lipid attachment site	Azaserine [HS] (436), sodium-*O*-vanadate [R] (40)
YGR226c	Protein of unknown function	Cycloheximide [HS] (50), diltiazem–HCl [S] (78), benzamidine [HS] (92), pH 4·16 [S] (703)
YGR221c	Protein of unknown function; similarity to hypothetical protein *YHR149c*	Guanosine-5-*O*-(2-thiodiphosphate) [S] (444)
YNL211c	Protein of unknown function	Sodium-*O*-vanadate [R] (40)
YGR231c/PHB2	Protein required for normal life-span; strong similarity to *S. cerevisiae PHB1p*, mammalian prohibitins and mouse B-cell receptor-associated protein BAP37 (Coates *et al.*, 1997)	Loperamide [S] (148)
YDL117w	Protein of unknown function; similarity to hypothetical *S. pombe* protein	Magnesium chloride [S] (17), 1,10 phenanthroline [S] (56), loperamide [HS] (148), azaserine [HS] (436)
YGR136w	Protein of unknown function; has similarity to *YPR154w*	EDTA [S] (2)
YBL047c	Protein with similarity to cytoskeletal protein *Uso1p*, *Pan1p* and mouse tyrosine kinase substrate EPS15; contains EF-hand calcium-binding domain and EH domain (Tang and Cai, 1996; Wendland and Emr, 1998)	Manganese chloride [HS] (23), diltiazem–HCl [S] (78), caffeine [HS] (484)
YDL074c	Protein of unknown function; weak similarity to spindle pole body protein *NUF1*	Cycloheximide [HS] (50), benomyl [R] (54), PCMB [S] (73), chlorpromazine [HS] (98), canthardate [S] (246), diphenyleneiodonium [S] (450)

Table 2. *Continued.*

	ORF/gene name: comments (MIPS/SGD/YPD)	Phenotypes (this study): hypersensitivity [HS]; sensitivity [S]; resistance [R]
YBR014c	Protein of unknown function; similarity to glutaredoxin and strong similarity to *YDL010w*	Diltiazem–HCl [S] (78)
YBR162c	Protein of unknown function; similarity to hypothetical protein *YJL171c* and *Aga1p*	EDTA [S] (2), copper sulphate [HS] (9)
YBR175w	Protein of unknown function; has WD (WD-40) repeats; weak similarity to GTP-binding proteins, *Tup1p*, *Pwp2p* and human LIS-1	Chlorpromazine [HS] (98), NDGA [S] (224), IPCPC [S] (598)
YJL051w	Protein of unknown function	EDTA [S] (2), cadmium chloride [HS] (6), benzamidine [HS] (92), valeryl salicylate [S] (146), 2,4 dinitrophenol [S] (156), sanguinarine [S] (160), phenyl-ethylamine [S] (282), 4-aminopyridine [HS] (390), hydroxyurea [S] (476), slow-growth (702), pH 3·8 [S] (703), N3 [HS] (704)
YJL059w/BTN1/YHC3	Protein of unknown function; similarity to human Batten disease-related protein CLN3, not required for mitochondrial function or degradation of mitochondrial ATP synthase (Pearce and Sherman, 1997)	Copper sulphate [HS] (9), loperamide [S] (148), extreme slow growth according to EUROSCARF [HS] (702)
YJL049w	Protein of unknown function	EDTA [S] (2)
YJL058c	Protein of unknown function; strong similarity to hypothetical protein *YBR270c*	Tungstic acid [HS] (4), copper sulphate [HS] (9), benomyl [HS] (54), diltiazem–HCl [S] (78), quercetin [S] (100), 6-dimethylaminopurine [S] (150), thiabendazole [HS] (310), polymyxin B [HS] (362), azaserine [S] (436), caffeine [HS] (484), pH 4·16 [HS] (703)
YJL056c/ZAP1	Metalloregulatory protein involved in zinc-responsive transcriptional regulation; plays a central role in zinc ion homeostasis by regulation transcription of the zinc uptake system genes in response to zinc; null mutant grows poorly on zinc-limiting media (Zhao and Eide, 1997)	EDTA [HS] (2), diltiazem–HCl [S] (78), loperamide [HS] (148), polymyxin B [S] (362), BAPTA [S] (438), IPCPC [S] (598)
YJL057c/IKS1	Probable serine/threonine kinase; null mutant is heat shock-sensitive; *ira1* kinase suppressor (Matviw *et al.*, 1993)	Copper sulphate [HS] (9)
YJL047c	Protein of unknown function; weak similarity to *cdc53p* and similarity to clathrin heavy chain in one domain	Thiabendazole [HS] (310), azaserine [HS] (436), IPCPC [HS] (598)
YJL055w	Protein of unknown function; similarity to *R. fascians* hypothetical protein 6; has similarity to *P. aeroginosa* hypothetical protein in azu region	Phenyl-ethylamine [R] (282), polymyxin B [R] (362)
YJL046w	Protein of unknown function; similarity to *E. coli* lipoate-protein ligase A	Zinc chloride [R] (24), hexadecylphosphocholine [R] (454)
YJL045w	Protein of unknown function; similarity to succinate dehydrogenase flavoprotein	Azaserine [HS] (436)

Table 2. *Continued.*

	ORF/gene name: comments (MIPS/SGD/YPD)	Phenotypes (this study): hypersensitivity [HS]; sensitivity [S]; resistance [R]
YJL065c	Protein of unknown function; weak similarity to DNA-directed DNA polymerase II chain C	2,2-Dipyridil [S] (44), diethyl maleate [R] (348)
YGL186c	Protein of unknown function; member of purine/cytosine permease family, a subfamily of the major facilitator superfamily (MFS) (Nelissen *et al.*, 1997); similarity to hypothetical protein *Fcy21p* and weak similarity to *FCY2* protein	Chlorpromazine [S] (98), 4-aminopyridine [HS] (390), hydroxyurea [HS] (476), slow-growth [S] (702), pH 4·16 [HS] (703)
YOL091w	Protein of unknown function; homozygous diploid null mutant fail to sporulate (Pearson *et al.*, 1998)	Grows much better than the wild-type under many different growth conditions, for example: 9, 24, 44, 56, 148, 300, 362, 436, 454, 701, etc.
YBR042c	Protein of unknown function; strong similarity to hypothetical protein *YDR018c*; probable ATPase with 4 potential transmembrane domains	Copper sulphate [R] (9)

stringent phenotypes (note that 'suggestive' phenotypes are not reported). In some cases we found that highly pleiotropic deletants displayed a slow growth phenotype, which could be observed already on glucose medium, YPD (e.g. YNL206c). However, this is not always the case. On the one hand, some extremely slow-growing deletants (like YJL059w) did not display a highly pleiotropic phenotype, being sensitive to only two compounds. On the other hand, some extremely pleiotropic deletants (like YOL018c) were not particularly slow growing on standard media. Finally, some hypersensitivity/resistance chemotypes could possibly result from changes in the cell membrane/wall permeability. However, in view of the large diversity of compounds used and chemotypes observed (see Tables 1 and 2) this explanation could not be true for the majority of chemotypes. It should be added that, in Table 2, only 'clear' phenotypes are listed. In addition to those mentioned, several 'suggestive' phenotypes (see Rieger *et al.*, 1997 for discussion of this point) have been observed. Since they were much weaker, reflect subtle variations in growth rate only, and frequently were different between the *MAT*α and *MAT*a series of deletants, their significance is doubtful and they have therefore not been presented in this study.

It remains to be established in a fully systematic manner that the chemotypes we observe co-segregate with the deletion; however, we have verified this in some cases. For example, co-segregation of the hypersensitivity to thiabendazole and IPCPC with the YJL047c and YNL206c deletions has been observed in tetrad analysis and was confirmed by complementation with the corresponding cognate wild-type genes. Furthermore, we have retested at random 12 compounds and nine mutants of opposite mating type (*MAT*a) from the EUROSCARF collection and in more than 80% of the cases the deletants were found to respond in a manner similar to that shown by the *MAT*α series of deletants.

The method used gives a large number of chemotypes, and we have found that over 50% of the deletants (which are listed in the MIPS/EUROSCARF database as 'normal' with no detectable phenotype) do have a significant growth deficiency in one or more of the conditions tested.

In conclusion, we suggest that the chemotyping approach described in this work is able to contribute to large-scale analysis of unknown gene function in several ways: (a) genes can be grouped into functional groups based on their observed sensitivity or resistance to different compounds; (b) important clues to gene function can be inferred

from a knowledge of the known pharmacological action of the agent which modifies growth of the corresponding mutant; and (c) generation of a scoreable phenotype from an otherwise silent mutation empowers a genetic approach to studying gene function. We conclude, therefore, that systematic chemotyping of yeast mutants will lead to a more precise understanding of the biochemical and physiological function for many of the 'functional orphan' yeast genes.

ACKNOWLEDGEMENTS

We gratefully acknowledge E. Sabire for engineering assistance, D. Church and A. Bernard for discussion and providing numerous compounds, and P. Koetter and K.-D. Entian from EUROSCARF (Johann Wolfgang Goethe-Universität Frankfurt, Institut für Mikrobiologie, Marie-Curie-Str. 9, Biozentrum N250, D-60439 Frankfurt am Main; e-mail: Euroscarf@em.uni-frankfurt.de) for providing the mutant strains used in this study and a collection of cloned wild-type genes. This work was supported by a grant from the E.C. (EUROFAN BIO4-CT95-0080. K.-J.R. received a fellowship from the E.C. (BIO4-CT95-0080).

REFERENCES

Alexis Biochemicals Catalog (1997/98). *Apoptosis, Cell Trafficking and Signal Transduction.* Alexis Corporation, Alte Hauensteinstrasse 4, CH-4448 Läufelingen, Switzerland.

Aronow, B., Toll, D., Patrick, J., Hollingsworth, P., McCartan, K. and Ullman, B. (1986). Expression of a novel high-affinity purine nucleobase transport function in mutant mammalian T lymphoblasts. *Mol. Cell. Biol.* **6**, 2957–2962.

Austriaco, N. R. Jr (1996). Review: to bud until death: the genetics of ageing in the yeast *Saccharomyces. Yeast* **12**, 623–630.

Baudin, A., Ozier-Kalogeropoulos, O., Denouel, A., Lacroute, F. and Cullin, C. (1993). A simple and efficient method for direct gene deletion in *Saccharomyces cerevisiae. Nucleic Acids Res.* **21**, 3329–3330.

Becker, M. A. and Kim, M. (1987). Regulation of purine synthesis *de novo* in human fibroblasts by purine nucleotides and phosphoribosylpyrophosphate. *J. Biol. Chem.* **262**, 14 531–14 537.

Bertani, L. E. and Campbell, J. L. (1994). The isolation and characterization of the gene (*dfr1*) encoding dihydrofolate reductase (DHFR) in *Schizosaccharomyces pombe. Gene* **147**, 131–135.

Boguslawski, G. (1992). *PBS2*, a yeast gene encoding a putative protein kinase, interacts with the *RAS2* pathway and affects osmotic sensitivity of *Saccharomyces cerevisiae. J. Gen. Microbiol.* **138**, 2425–2432.

Bonangelino, C. J., Catlett, N. L. and Weisman, L. S. (1997). *Vac7p*, a novel vacuolar protein, is required for normal vacuole inheritance and morphology. *Mol. Cell. Biol.* **17**, 6847–6858.

Calvert, C. M. and Sanders, D. (1995). Inositol trisphosphate-dependent and -independent Ca^{2+} mobilization pathways at the vacuolar membrane of *Candida albicans. J. Biol. Chem.* **270**, 7272–7280.

Castano, I. B., Heath-Pagliuso, S., Sadoff, B. U., Fitzhugh, D. J. and Christman, M. F. (1996). A novel family of TRF (DNA topoisomerase I-related function) genes required for proper nuclear segregation. *Nucleic Acids Res.* **24**, 2404–2410.

Chetelat, A., Albertini, S., Dresp, J. H., Strobel, R. and Gocke, E. (1993). Photomutagenesis test development: I. 8-Methoxysporalen, chlorpromazine and sunscreen compounds in bacterial and yeast assays. *Mutat. Res.* **292**, 241–250.

Chong, J. P. J., Mahbubani, H. M., Khoo, C.-Y. and Blow, J. J. (1995). Purification of an MCM-containing complex as a component of the DNA replication licensing system. *Nature* **375**, 418–421.

Coates, P. J., Jamieson, D. J., Smart, K., Prescott, A. R. and Hall, P. A. (1997). The prohibitin family of mitochondrial proteins regulate replicative lifespan. *Curr. Biol.* **7**, 607–610.

De Risi, J. L., Iyer, V. R. and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686.

Du, L. L., Collins, R. N. and Novick, P. J. (1998). Identification of a *Sec4p* GTPase-activating protein (GAP) as a novel member of a Rab GAP family. *J. Biol. Chem.* **273**, 3253–3256.

Felix, M. A., Pines, J., Hunt, T. and Karsenti, E. (1989). Temporal regulation of *cdc2* mitotic kinase activity and cyclin degradation in cell-free extracts of Xenopus eggs. *J. Cell Sci. Suppl.* **12**, 99–116.

Goffeau, A., Park, J., Paulsen, I. T., Jonniaux, J. L., Dinh, T., Mordant, P. and Saier, M. H. Jr (1997). Multidrug-resistant transport proteins in yeast: complete inventory and phylogenetic characterization of yeast open reading frames with the major facilitator superfamily. *Yeast* **13**, 43–54.

Holthuis, J. C. M., Nichols, B. J., Dhruvakumar, S. and Pelham, H. R. B. (1998). Two syntaxin homologues in the TGN/endosomal system of yeast. *EMBO J.* **17**, 113–126.

Huhse, B., Rehling, P., Albertini, M., Meller, K. and Kunau, W. H. (1998). *Pex17p* of *Saccharomyces cerevisiae* is a novel peroxin and component of the peroxisomal protein translocation machinery. *J. Cell Biol.* **140**, 49–60.

Jensen, E. C., Ogg, C. and Nickerson, K. W. (1992). Lipoxygenase inhibitors shift the yeast/mycelium dimorphism in *Ceratocystis ulmi. Appl. Environ. Microbiol.* **58**, 2505–2508.

Johnson, J. L., Rajagopalan, K. V. and Cohen, H. J. (1974). Molecular basis of the biological function of molybdenum. *J. Biol. Chem.* **249**, 859–866.

Kuge, S., Jones, N. and Nomoto, A. (1997). Regulation of yAP-1 nuclear localization in response to oxidative stress. *EMBO J.* **16**, 1710–1720.

Kule, C., Ondrejickova, O. and Verner, K. (1994). Doxorubicin, daunorubicin, and mitoxantrone cytotoxicity in yeast. *Mol. Pharmacol.* **46**, 1234–1240.

Kustimur, S., Memis, L., Kilinc, M. and Ercan, Z. S. (1997). Effect of nordihydroguaretic acid and fluconazole on the LTC4/PGE2 ratio in the kidney of mice damaged by *Candida albicans*. *Prostaglandins Leukot. Essent. Fatty Acids* **56**, 281–283.

Lesuisse, E., Casteras-Simon, M. and Labbe, P. (1996). Evidence for the *Saccharomyces cerevisiae* ferrireductase system being a multicomponent electron transport chain. *J. Biol. Chem.* **271**, 13 578–13 583.

Loukin, S. and Kung, C. (1995). Manganese effectively supports yeast cell-cycle progression in place of calcium. *J. Cell. Biol.* **131**, 1025–1037.

Matviw, H., Yu, G. and Young, D. (1993). Identification and genetic analysis of *Schizosaccharomyces pombe* cDNAs that suppress deletion of *IRA1* in *Saccharomyces cerevisiae*. *Gene* **129**, 147–152.

Morino, Y., Kojima, H. and Tanase, S. (1979). Affinity labeling of alanine aminotransferase by 3-chloro-L-alanine. *J. Biol. Chem.* **254**, 279–285.

Nelissen, B., De Wachter, R. and Goffeau, A. (1997). Classification of all putative permeases of the major facilitator superfamily encoded by the complete genome of *Saccharomyces cerevisiae*. *FEMS Microbiol. Rev.* **21**, 113–134.

Patel, S., Sprung, A. U., Keller, B. A., Heaton, V. J. and Fisher, L. M. (1997). Identification of yeast DNA topoisomerase II mutants resistant to the antitumor drug doxorubicin: implications for the mechanisms of doxorubicin action and cytotoxicity. *Mol. Pharmacol.* **52**, 658–666.

Pearce, D. A. and Sherman, F. (1997). *BTN1*, a yeast gene corresponding to the human gene responsible for Batten's disease, is not essential for viability, mitochondrial function, or degradation of mitochondrial ATP synthase. *Yeast* **13**, 691–697.

Pearson, B. M., Hernando, Y. and Schweizer, M. (1998). Construction of PCR-ligated long flanking homology cassettes for use in the functional analysis of six unknown open reading frames from the left and right arms of *Saccharomyces cerevisiae* chromosome XV. *Yeast* **14**, 391–399.

Poli, F., Pancaldi, S., Dall'Olio, G. and Fasulo, M. P. (1995). Morphogenetic effects induced by spermine and ruthenium red in yeasts. *Cytobios.* **81**, 201–211.

Rieger, K.-J., Kaniak, A., Coppée, J.-Y., Aljinovic, G., Baudin-Baillieu, A., Orlowska, G., Gromadka, R., Groudinsky, O., di Rago, J.-P. and Slonimski, P. P. (1997). Large-scale phenotypic analysis—the pilot project on yeast chromosome III. *Yeast* **13**, 1547–1562.

Rieger, K.-J., Orlowska, G., Kaniak, A., Coppée, J.-Y., Aljinovic, G. and Slonimski, P. P. (1998). Large-scale phenotypic analysis in microtiter plates of mutants with deleted open reading frames from yeast chromosome III: key-step between genomic sequencing and protein function. *Methods in Microbiology*. Automation, Academic Press (in press).

Ruiz, T. and Todriguez, L. (1986). Effect of antical-modulin drugs on the action of yeast alpha factor pheromone. *Arch. Microbiol.* **145**, 104–106.

Tachikawa, H., Funahashi, W., Takeuchi, Y., Nakanishi, H., Nishihara, R., Katoh, S., Gao, X. D., Mizunaga, T. and Fujimoto, D. (1997). Overproduction of *Mpd2p* suppresses the lethality of protein disulfide isomerase depletion in a CXXC sequence dependent manner. *Biochem. Biophys. Res. Commun.* **239**, 710–714.

Tang, H. Y. and Cai, M. (1996). The EH-domain-containing protein *Pan1* is required for normal organization of the actin cytoskeleton in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **16**, 4897–4914.

Verna, J., Lodder, A., Lee, K., Vagts, A. and Ballester, R. (1997). A family of genes required for maintenance of cell wall integrity and for the stress response in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **94**, 13 804–13 809.

Wendland, B. and Emr, S. D. (1998). *Pan1p*, yeast eps15, functions as a multivalent adaptor that coordinates protein–protein interactions essential for endocytosis. *J. Cell. Biol.* **141**, 71–84.

Yu, P. H. (1994). Pharmacological and clinical implications of MAO-B inhibitors. *Gen. Pharmacol.* **25**, 1527–1539.

Zhao, H. and Eide, D. J. (1997). *Zap1p*, a metalloregulatory protein involved in zinc-responsive transcriptional regulation in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **17**, 5044–5052.

Zumstein, E., Pearson, B. M., Kalogeropoulos, A. and Schweizer, M. (1995). A 29·425 kb segment on the left arm of yeast chromosome XV contains more than twice as many unknown as known open reading frames. *Yeast* **11**, 975–986.

GENOME ANALYSIS

A LABORATORY MANUAL

VOLUME 3

CLONING SYSTEMS

Volume Editors

Bruce Birren

Eric D. Green

Sue Klapholz

Richard M. Myers

Harold Riethman

Jane Roskams

COLD SPRING HARBOR LABORATORY PRESS

GENOME ANALYSIS

A LABORATORY MANUAL
VOLUME 3 / CLONING SYSTEMS

Printed in the United States of America
Book design by Emily Harste
Cover design by Tony Urgo

Library of Congress Cataloging-in-Publication Data

Genome analysis: a laboratory manual / editors Eric D. Green ... [et al.]
v. <1-2 >. cm.
Includes bibliographical references and index
Contents: v. 1. Analyzing DNA ISBN 0-87969-496-3 (comb). ISBN 0-87969-495-5 (cloth). v. 2. Detecting genes ISBN 0-87969-511-0 (comb). ISBN 0-87969-510-2 (cloth). v. 3. Cloning Systems ISBN 0-87969-513-7 (comb). ISBN 0-87969-512-9 (cloth). v. 4. Mapping Genomes ISBN 0-87969-515-3 (comb). ISBN 0-87969-514-5 (cloth).
1. Gene mapping -- Laboratory manuals. 2. Nucleotide sequence -- Laboratory manuals. I. Green, Eric D.
QH445.2.G4645 1997
576.5'3—dc21

97-17117
CIP

Cover caption for paperback
Top: Photographs of human infant and mouse were reproduced with permission from Comstock, Inc.
Middle: (*Left*) *S. cerevisiae;* (*center*) *C. elegans;* (*right*) *Drosophila*
Bottom: E. coli

All Cold Spring Harbor Laboratory Press publications may be ordered directly from Cold Spring Harbor Laboratory Press, 10 Skyline Drive, Plainview, New York 11803-2500. Phone: 1-800-843-4388 in Continental U.S. and Canada. All other locations: (516) 349-1946. E-mail: cshpress@cshl.org. For a complete catalog of all Cold Spring Harbor Laboratory Press publications, visit our World Wide Web site: http://www.cshl.org.

PROTOCOL

Robotic Replication

1. Label microtiter plates and fill with liquid medium as described in steps 1 and 2 of the Manual Replication protocol (see p. 18).

   Note: The labeling of plates and addition of medium can be performed in a separate area, either in a laminar flow hood, if one is available, or on an open bench area using standard aseptic techniques.

2. Set up the robot as follows:
   a. Fill the sterilization bath with 95% ethanol to a depth greater than the depth of culture in the microtiter plates.
   b. Fill the sonication bath with sterile distilled H_2O containing 0.1% (v/v) decon.
   c. Switch on the air filter.
   d. Attach the appropriate replicating tool (either a 96-pin or 384-pin tool) to the moving head.
   e. Clean the pins of the replicating tool by placing it in the sonication bath. Sonicate for 3 minutes.
   f. Sterilize the pins by immersion in the ethanol bath for approximately 15 seconds. Air-dry for approximately 10 seconds.

3. Place the source and recipient plates on the bed of the robot. Remove the plate lids just before beginning the replication procedure.

4. Set up a replicating routine as follows:
   a. Immerse the pins of the replicating tool in the sterilization bath for approximately 15 seconds.
   b. Air-dry for approximately 10 seconds.
   c. Immerse the pins in the wells of the first source plate. To avoid any splashing of culture, slowly remove the tool from the plate.
   d. Immerse the pins in the wells of the first recipient plate and remove the tool.

   Note: The thickness of the replicating pins determines the volume of culture to be transferred and therefore influences the rate of growth in the recipient plate. Multiple inoculations may be required to achieve sufficient growth at 37°C in 16–20 hours, and this should be tested empirically with a small number of plates.

5. Repeat step 4(a–d) with subsequent plates.

6. Replace the lids as soon as inoculation of the plate is completed. Freeze the source plates as described in step 6 (see p. 17) and incubate recipient plates at 37°C overnight as described in step 5 (see p. 17).

PROTOCOL

Recovery of Individual Clones

The following procedure should be performed in a laminar flow hood if one is available. Otherwise, perform the procedure on an open bench area using standard aseptic techniques.

1. Prepare a list of the plates and addresses of the clones to be recovered.

   Note: Organizing this list in accordance with the system used to store the plates in the freezer will speed the process of plate retrieval and reduce the amount of time freezer doors must be kept open.

2. Withdraw only the desired plates from storage at –80°C and place them in a tray on a layer of dry ice.

3. Remove the plate lid. Blot any liquid that may have condensed on the surface of the plate with sterile, absorbent paper.

   Note: It is important to blot any liquid from the surface of the plate to reduce well-to-well contamination in the library, particularly for library copies in 384-well plates (where there is no physical barrier to liquid spread by capillarity). Use sterile Whatman 3MM for this purpose (sterilized by autoclaving).

4. Pick a small amount of the frozen culture with a sterile wooden cocktail stick (toothpick) or wooden applicator stick. Streak the culture onto an LB-agar plate containing the appropriate antibiotic.

   Notes: For appropriate concentrations of antibiotics in culture medium and the preparation of stock and working solutions, see Appendix.

   When picking, use a drilling motion with the stick, rather than a levering motion which may result in frozen culture being transferred to neighboring wells.

   Some laboratories use a sterile mask with a single-well-size hole (e.g., made from Whatman benchkote paper) to avoid cross-contamination between cultures in separate wells.

5. Replace the plate lid and return the plate to its stack in the freezer. Incubate the agar plates at 37°C overnight to allow growth of colonies.

Library Screening

Several effective strategies for screening genomic libraries for clones of interest have been developed. These are based on colony hybridization, PCR, or a combination of both techniques. Colony hybridization with bacterial clones is extremely robust and efficient, typically yielding clear, strong signals even when the cloned DNA is present as a single copy per cell (e.g., BACs and fosmids; Kim et al. 1994). An advantage of screening by hybridization is the variety of different types of probes that can be used, including those corresponding to single-copy sequences, repetitive elements, insert end segments (e.g., for "walking" efforts), and whole clone inserts (e.g., purified YAC DNA, BAC DNA, or PCR products). The routine of arraying and storing genomic libraries in microtiter plates facilitates the production of filters containing colonies at varying densities. When robotics are used, colonies can be arrayed at a very high density in a highly reproducible fashion. Approximately 20,000 clones can be represented on a 22 x 22-cm filter, and an 8 x 12-cm filter can hold several thousand clones. Alternatively, pools of DNA derived from well-defined combinations of microtiter plates or portions thereof can be generated from arrayed libraries and used for PCR-based screening. Such pools can be used effectively for rapid and efficient screening of the library using specific PCR assays. One drawback to PCR-based screening, however, is that the amount of effort required for generating suitable DNA pools goes up dramatically as the number of clones in a library increases. This is particularly the case if the screening strategy is designed to allow the identification of individual positive clones (as opposed to a delimited set of clones, such as an individual microtiter plate). As a result, a "hybrid" screening strategy using a combination of PCR and colony hybridization can be used as a compromise for genomic library screening. Specifically, PCR screening is used to identify individual microtiter plates containing positive clones. Then, colony hybridization is performed with those microtiter plates to identify the precise well positions of positive clones (Green and Olson 1990a). However, although in the majority of cases the clones identified by hybridization of a PCR product will be the same as those detected by PCR, the specificity of an STS as a hybridization probe is not always identical to the specificity of its corresponding PCR assay (e.g., due to the presence of repetitive sequences residing between the two PCR primers).

A great majority of investigators perform hybridization- or PCR-based screening of genomic libraries using hybridization filters or DNA pools obtained from central facilities or companies (see discussion on pp. 2–3). However, it may on occasion be necessary to produce filters for hybridization screening or DNA pools for PCR screening in a local laboratory. Thus, appropriate protocols are provided for each of these in the following sections.

Meeting report

Screensavers: trends in high-throughput analysis

Even, and perhaps especially, in the post-genomic era, the search for new drugs begins with the detection and (hopefully) validation of a novel target, and with the development of an assay for the interaction of an artificial agonist or antagonist – and of its natural ligand, if known – with that target.

As the race to develop new chemical entities with novel pharmacological activity hots up, the major pharmaceutical companies have invested heavily in automation and robotics, and in the high-throughput assaying of large compound libraries for pharmacological activities against drug receptors. The current state of the art in this area was the theme of a recent international conference*, which reflected the intense applied interest in this area by the fact that some 98% of the 1256 delegates were from industrial corporations.

Cruising the post-genomic orphanage

Both the results of the whole-genome-sequencing programmes and the contents of the extensive expressed sequence tag (EST) databases have made it abundantly clear that many gene products have no close relatives in the databases (they are 'orphans') and are of unknown function, that many are cell-membrane receptors (especially G-protein-coupled receptors), that their natural ligands are also unknown, and that they consequently represent important novel targets for both agonists and antagonists.

The newly discovered orexins A and B and their receptors are a case in point (Masashi Yanagisawa, University of Texas Southwestern Medical Center, Dallas, TX, USA), and are involved in the control of food intake. Direct binding assays are possible but measurements based on the agonist-dependent production of Ca^{2+} transients in transfected cells provide a more pertinent, functional assay. It is recognized that there are some 10 000–40 000 different mRNAs in the 500 000 molecules expressed in a typical mammalian cell. This means that, of the many novel approaches to functional genomics, expression profiling of these mRNAs using oligonucleotide arrays has enormous power for providing clues to the function of orphan genes (Eugene Brown, Genetics Institute, Cambridge, MA, USA). Modern systems of this type are linear from 0.5 to 500 pM target RNA and are reproducible and quantitative. Such arrays, preferably prepared using bacterial artificial chromosomes, allow the facile detection of huge numbers of single-nucleotide polymorphisms (Janice Kurth, Genset Corporation, La Jolla, CA, USA). These can be exploited in screening patients for their likelihood of acquiring particular illnesses and for their suitability for chronic drug therapy.

What you see is what you get – novel optical methods for high-throughput screening

Fluorescent methods are probably the methods of choice for high-throughput screening (HTS) assays, and their repertoire continues to increase. Those based on variants of time-resolved fluorescence (which allows the discrimination of the signal of interest from other fluorescent background signals) have particular merit, especially as the drive towards miniaturization means that, because of contributions from the assay plates, the signal decreases more quickly than the background as the assay volume is reduced (Jack Owicki, LJL Biosystems, Sunnyvale, CA, USA).

The current move is away from the traditional 96-well plate to 384- and especially 1536-well versions, where reagent costs are typically 100 times lower and assay volumes drop from some 400 μl to 5–10 μl (Jonathan Burbaum, Pharmacopoeia, Princeton, NJ, USA). Technical issues become significant here, such as the use of conical rather than square wells to avoid wicking and the importance of measures to stop evaporation, but the great benefit is cost reduction, with typical costs for a screening campaign being reduced from US$35 million to US$1.1 million. Best of all is if there are no reagents. The use of infrared spectroscopy in HTS is a novel, reagentless and generic technique requiring at most a few μl of sample; as an example from titre-improvement programmes, the measurement time may be reduced to 1 sec from the 15 min required for the traditional HPLC analysis (Douglas Kell, University of Wales, Aberystwyth, UK).

Classical analysis of optical assays in microtitre plates used scanning methods in which the results were read sequentially by a single detector. This represented a substantial bottleneck in the speed of the overall screening process and thus a major trend is towards imaging methods in which, by coupling a telecentric (non-parallax) lens and a CCD camera, an entire plate may be imaged and read simultaneously (Ronald Barrett, Affymax Research Institute, Palo Alto, CA, USA; Neil Cook, Amersham Pharmacia Biotech, Cardiff, UK). To achieve these levels of sensitivity (at which the photon flux may be a hundredth to a ten-thousandth of that of starlight), improvements are required in all areas, with reagents, hardware and software all contributing to the achievement of the required sensitivity.

Fibre-optic arrays provide another means of interrogating many assays in parallel; etching microwells onto the end of such optical fibres allows assays to be performed in volumes as low as 90 fl (David Walt, Tufts University, North Grafton, MA, USA). More accessibly, confocal methods exploiting fluorescence-correlation spectroscopy (FCS) can interrogate a volume of 1 μm^3 (i.e. 1 fl), in which a 10 nM solution of a fluorophore contains on average six molecules. Analysis of the time course of fluctuations in their number density provide much information on their molecular environment and, in particular, on whether they are bound or free; any 'traditional' fluorescence assay may be configured for FCS (Keith Moore, SmithKline Beecham, Harlow, UK).

The numbers game; tracking chemical diversity

Imagine that there are just ten crucial and independent parameters ('explanatory variables') that can contribute to a drug's activity and

*The 4th Annual Meeting of the Society for Biomolecular Screening (http://www.sbsonline.org/) was held in Baltimore, MD, USA, 20–24 September, 1998.

that obtaining a lead compound with a binding constant of 1 μM or better requires that each of them is within ±10% on a linear scale of the 'correct' or optimal value. This means that we are looking for a single entity in 5^{10} possible sets of properties (and for a precision of just twice this per explanatory variable, there are 10^{10} possibilities). Considerations such as this have led to the explosion of libraries of candidate drug compounds, typically containing 10^5–10^6 pure substances. To produce these is therefore as significant as the need to analyse their pharmacological behaviour, and to assess and maximize the chemical diversity within a library is of particular importance.

However, using the methods of combinatorial chemistry combined with evolutionary algorithms means that a potential library of 160 000 (from a four-step chemical synthesis in which there are 10, 10, 40 and 40 reagents of each type) can be decreased to just 400 experiments in which after a 'random' 20 mixtures, each reagent type is optimized over 19 further generations (Klaus Gubernator, Combichem, San Diego, CA, USA). Indeed, the combination of chemical diversity with intelligent computer analysis is crucial to this type of enterprise. Both active and inactive compounds contribute useful information as one seeks to decrease the search space for useful pharmacophores, especially if structural (rather than biophysical) parameters are used as the inputs (Susan Bassett, Bioreason, Santa Fe, NM, USA). In terms of finding metrics for assessing diversity (a generic problem that may also be applied to biodiversity metrics), good metrics for library design must always cluster molecules with similar biological activities together (Dora Schnur, Pharmacopeia, Princeton, NJ, USA), and yet diversity metrics should seek to minimize the proximity of molecules in descriptor space (Adrienne Tymiak, Bristol-Myers Squibb Pharmaceutical Research Institute, Buffalo, NY, USA). A similarity index based on the average number of shared atom-pairing descriptors over the total number of such descriptors provides a robust metric for such analyses.

Both chemical libraries and the results of assays using them produce huge amounts of data, which many workers in a large organization may wish to access. The design and construction of appropriate databases is thus another great need, and the interface must be constructed in a way that both makes navigation easy for the bench scientist and permits the facile deployment of powerful query and report tools. Only hierarchical methods permit this with any convenience, and allow the user to organize data from thousands of drug screens. The *Discovery explorer* tool is one such implementation, which provides decision support via a scalable, robust, flexible and enterprise-wide architecture (Anthony Kreamer, SmithKline Beecham, King of Prussia, PA, USA).

Given that the human genome probably contains some 70 000 genes, that one might wish to assay some 10% of these and that half may be amenable to direct binding assays, the big pharmaceutical companies will certainly need to be looking at the results of several thousand screens (Mario Geysen, GlaxoWellcome, Research Triangle Park, NC, USA). If combinatorial chemistry is to be the main source of leads (as well as natural products), only the split-and-mix strategy of synthesis on solid supports (beads) is appropriate; discrimination between beads may be carried out by encoding them via a linker labelled with stable isotopes in various ratios.

Small is beautiful: miniaturization in ultra-HTS

A common, if arbitrary, definition of a system for ultra-HTS is one in which 100 000 assays are run per day. This is slightly more than one per second and requires careful integration of the necessarily robotic systems, which deploy compound libraries, run the assays and analyse the data. With primary hit rates typically running at 0.1%, it is critical to minimize both false positives and false negatives, and to ensure that the miniaturized assays in 1536-well plates with volumes under 10 μl behave exactly like those carried out in the test tube or the 96-well plate. Even the 1536-well plate has its competitors, as laboratory-on-a-chip systems (in which reagents, cells and drug candidates are mixed by electrokinetic forces operating in microfluidic channels of 10–100 μm) allow complex assays such as those for Ca^{2+} release to be operated at rates of 2000 cells min^{-1} and allow several thousand replicates to be analysed in a total volume of less than 20 μl (Michael Knapp, Caliper Technologies, Palo Alto, CA, USA).

However, analyses done under these ultra-HTS conditions must be considerably more robust than those initially developed by the scientists studying novel targets. The optimization of such assays represents a combinatorial problem as intractable as that described above regarding the statistical difficulty of optimizing a drug lead [even the question of whether to include one of 16 buffer components – never mind optimizing its concentration – gives 2^{16} (or 65 536) possibilities requiring 683 96-well plates if all are to be assessed]. Modern methods of experimental design can reduce this to just two or three such plates, and their preparation may be integrated into a laboratory robotics system (Frances Stewart, SmithKline Beecham, King of Prussia, PA, USA).

As assays become smaller, we enter the field of nanotechnology; nm-sized gold microspheres possess unusual optical properties (such as a molar extinction coefficient of some 10^9 M^{-1} cm^{-1}) that change dramatically upon ligand binding. They have great potential for exploitation in different types of binding measurements based on surface-plasmon resonance (Michael Natan, Pennsylvania State University, University Park, PA, USA) and in the highly selective detection of DNA. In this technique, the chromophoric changes attending nucleic-acid binding to gold microspheres derivatized with complementary nucleic acids are both unusually temperature dependent and permanent (Chad Mirkin, Northwestern University, Evanston, IL, USA).

New ways to analyse cellular properties may also be greatly assisted by modifying the biology. The yeast two-hybrid system is a well-known method of detecting protein–protein interactions *in vivo* but traditional versions can be rather tedious, as cells need to be cotransfected with both putative partners of the binding event. A new variant has, however, been developed in which an entire cDNA library containing the 'prey' is held in one yeast mating type and cells of the other mating type are transformed with the 'bait' (Yiwu He, GlaxoWellcome, Research Triangle Park, NC, USA). Coincubation of the cells followed by a chemiluminescent β-galactosidase reporter assay

means that one person can supervise the automated performance of 50 full assays per month, each typically producing 20 binding partners that might provide novel drug targets.

Interrogating single molecules is clearly the ultimate in miniaturization and, although such methods cannot yet easily be parallelized, scanning-probe methods such as atomic-force microscopy with functionalized probes allow the direct and elegant measurement and discrimination of the interaction between single molecules (Saul Tendler, Nottingham University, Nottingham, UK).

And was it all worth it?

Is the modern marriage of biochemical-, genomic- (and intuition-) based target development and HTS leading to new and useful drugs? Two examples suffice to indicate that the answer is resoundingly in the affirmative.

Viramune is a novel, nonnucleosidic drug active against the reverse-transcriptase enzyme of the human-immunodeficiency-virus 1 (the major causative agent of AIDS), and was approved for use in 1996 (John Proudfoot, Boehringer Ingelheim Pharmaceuticals, Ridgefield, CT, USA). In 1988, it was known only that the target enzyme contained two subunits (an X-ray structure became available in 1992) and a drug screen against it was initiated (at this time, HTS meant 100–200 compounds per week!). After testing approximately 1600 molecules, of which 1% gave some kind of 'hit', a lead compound was discovered with an IC_{50} of 6 μM. Optimization of this compound led to a novel chemical entity with excellent pharmacological properties and an IC_{50} of 35 nM, low enough that therapeutic doses (giving a blood concentration of some 25 μM) were active even against 'resistant' strains.

It is now well established that the many complications of diabetes, such as nephropathy, neuropathy and, in particular, retinopathy leading to blindness, are exacerbated when glucose levels are not well controlled. What has been much less clear is the mechanism by which chronic hyperglycemia actually causes these complications. Following detailed work at Sphinx Pharmaceuticals, which developed assays specific for the many kinds of protein kinase, it was hypothesized that it was, in fact, an excessive activity of the β isoforms of protein-kinase C (PKC) induced by the higher levels of diacylglycerol formed under conditions of glucose excess that might be the major culprit (William Heath, Lilly Research Laboratories, Indianapolis, IN, USA). Despite management opposition (because non-selective inhibitors of the PKC family were known to be highly cytotoxic), a series of screens was developed against each of the eight human PKC isozymes. An indolecarbazole natural product related to staurosporine led to the development of LY 333531, a highly selective molecule with an IC_{50} of 5 nM against the β isoforms but several hundred nM against the rest. Interestingly, even though the screen was intended to find molecules acting on the regulatory site of the PKC enzyme, it was actually the catalytic (ATP-binding) site that is the target, and the competitive inhibition was only observed when the assay was run at ATP levels significantly lower than those thought to be prevalent *in vivo*! 'The ATP level in the cell is 1 mM, but not everywhere in the cell.'

Assay I say

In conclusion, it is evident that the trend towards miniaturization, the intelligent generation and deployment of chemical libraries, the innovative hardware and software, and the robust automation now available are major forces in the drive to develop new pharmaceuticals with novel targets, high efficacy and, of course, substantial commercial potential. However, for these hopes to be realized, good assays will remain paramount.

Douglas Kell
Institute of Biological Sciences, Aberystwyth, UK SY23 3DD.
(E-mail: dbk@aber.ac.uk;
WWW: http://gepasi.dbs.aber.ac.uk/home.htm)

The sheep and the goats

Meeting report

Communication between the makers and the users of laboratory equipment is always a problem. The user wants to know what is available, and to tell the maker the problems with their equipment; the maker wants to make sure that the user knows about the latest products, and to make sure their own products are as good as they can be. A recent meeting* gave an excellent opportunity for this sort of exchange of information.

The speakers were a varied mixture – some from the manufacturers of different types of separation equipment, extolling the virtues of their latest devices, others from the sharp end, reporting their own experience of various techniques, yet others discussing issues arising from the procedures involved.

Products

There are many different sorts of separation equipment out there, including classical chromatography columns, packed and fluidized beds, and membrane-based systems. Harvey Brandwin (Pall Filtron, Northborough, MA, USA) discussed some of the problems of membrane-based systems, including the need to optimize each system individually, but emphasized their advantages, especially in virus removal. He also reported that Pall Filtron have developed new filters capable of 3-log removal of viruses down to 20 nm. When using filters with such narrow pores, it is essential to use prefilters, to prevent the filter rapidly clogging up, and they will also help when using larger-pore-size filters.

W-D. Schleuning (Schering, Berlin, Germany) introduced new surrogate systems for evaluating the biological activity of new agents to replace the standard animal models. These included the use of hormone-sensitive promoters in yeast and the two-hybrid system, and also touched on issues of high-throughput-screening systems, which are already moving from 96-well to 384-well plates, and may in the future move to 864- and even 1536-well plates.

**Bio-Europe '98: Bioseparation and Bioprocessing of Biological Molecules*, the eighth annual meeting organized by G. Subramanian, was held in Cambridge, UK, 7–9 September 1998.

Drug screening—beyond the bottleneck

High-throughput screening still has a long way to go if it is to achieve its often-touted potential.

Alan Dove

Screening for a drug is often likened to searching for a needle in a haystack. In fact, with the application of combinatorial approaches and parallel synthesis to discovery, the sheer scale and diversity of chemical libraries have made identifying drug leads more akin to searching for a needle in many different haystacks. Today, screening programs may process up to 100,000 compounds a day, a thousand times as many as were processed in an entire week in 1990. To complicate the problem still further, a surfeit of new drug targets and their variants is emerging from genomics efforts.

Faced with the daunting prospect of screening thousands, or even hundreds of thousands, of chemical structures in one day, against an expanding universe of targets, big pharma has turned to biotechnology companies that specialize in equipment and services for high-throughput screening (HTS)—a term applied to several types of technologies for rapid testing or development of pharmaceutical lead compounds. The majority of these companies offer a range of products and strategies that are as varied as they are numerous. Whether they can deliver results that match their ambitious claims remains uncertain. Already, pharmaceutical companies are rapidly learning that clearing one obstacle in drug development often brings another into view. In fact, accomplishment of drug development goals may require a change not only in technology infrastructure, but also in the culture of pharmaceutical research itself.

Increasing throughput

The systematic screening of natural compounds (and more recently synthetic ones) to find potential drug leads has been a cornerstone of pharmaceutical research for many decades. Typically, a screening program commences with the development of an assay—originally performed in vivo, but now almost exclusively done using cells or targets in vitro—intended to mimic some aspect of a disease of interest. Assay development is then carried out to determine optimal conditions and ensure that the system under study is both robust and reproducible. With a hit rate of 0.1–0.5%, this is particularly important, in that false positives and negatives can lead to squandered time or resources in the pursuit of unsuitable leads.

The move to increase throughput (i.e., the number of individual drug activity assays performed in a given time) has focused on two areas. First, test volumes are shrinking from milliliters to microliters (and with miniaturization even nanoliters; see "Cutting screening down to size"). Thus, the 96-well microplate replaced 1 ml assays in the 1980s, the 384-well plate was introduced in 1994, and the 1,536-well plate (with a capacity of 1–2 μl) is now being touted as the new industry standard.

Second, processes are being automated to make them less labor intensive, more reproducible, and less expensive. Robots now perform repetitive tasks like pipetting, assay reading, or sample storage, allowing compounds to be retrieved from an automated dispensary, applied to cultured cells for assay under tightly controlled conditions, and scored for activity without human intervention. Some companies have begun to refer to these new technologies, which can push screening rates as high as 100,000 assays per day (slightly more than one per second), as ultrahigh-throughput screening (UHTS)[1].

In solving the problem of assay replication, however, the automated systems have uncovered the difficulty of integrating diverse pieces of machinery and incorporating the mass production operation into traditional R&D culture. "As a rule, most of the major companies are buying what are generally referred to as platform technologies. They've gone out and bought what they feel are the best technologies in each area," explains Richard Archer, CEO of The Automation Partnership (Cambridge, UK). The problem is that the resulting collection of machinery may have to be reengineered to work as an integrated system. Archer says

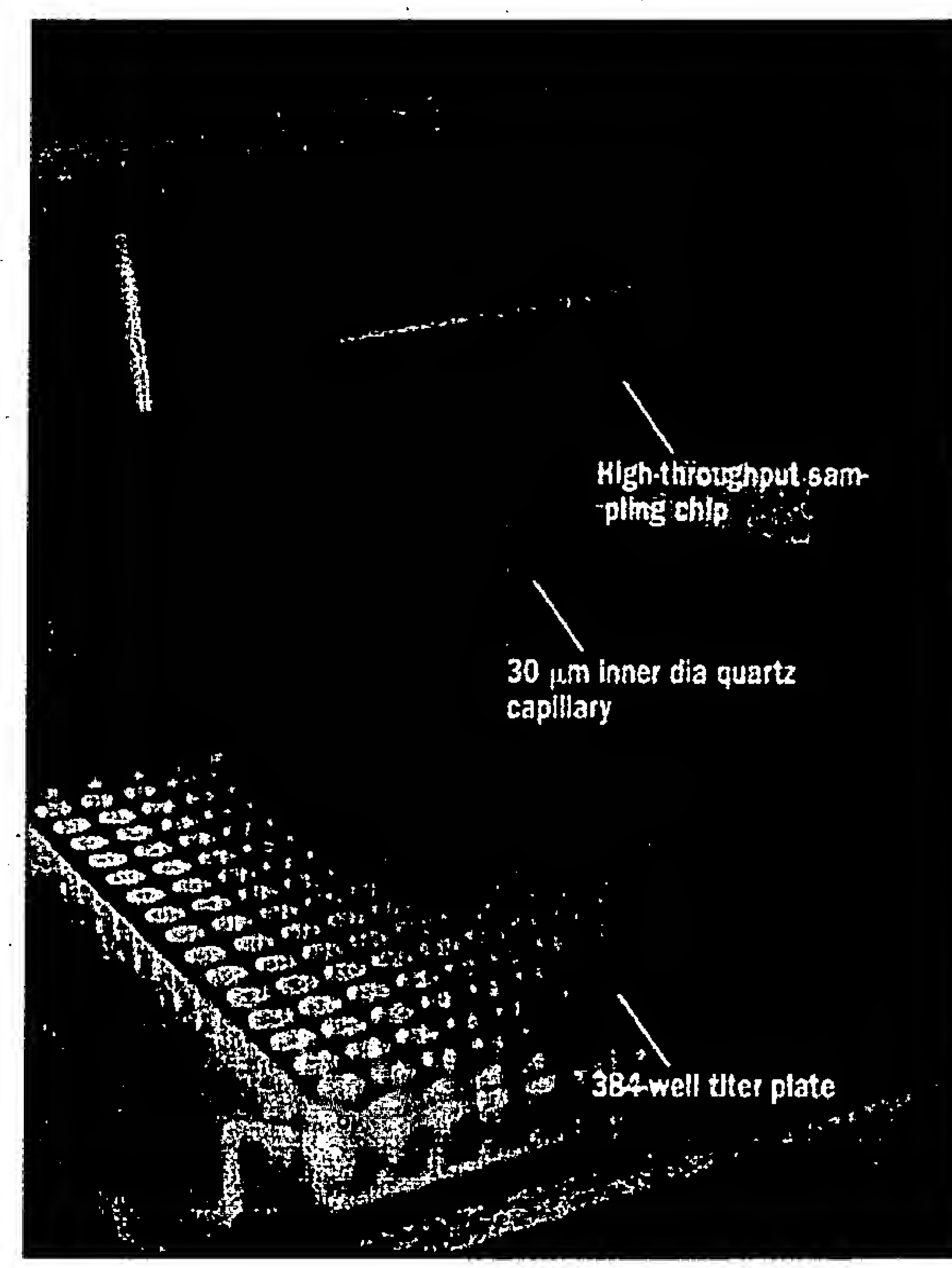


Figure 1. Caliper's microfluidic chips are now being refined for use in HTS.

his company is working to solve this problem by marketing "big boxes"—complete general-purpose HTS systems that can be purchased outright, providing everything from sample storage and automated cell culture to assay readout.

Taking the integrated system concept a step further, Aurora Biosciences (San Diego, CA) sells its HTS systems in conjunction with proprietary cellular assays (see below). The integration of HTS and assay comes at a price, however. Rather than selling its systems off the shelf, Aurora licenses its technology with reach-through agreements that include milestone payments and royalties on any drugs developed

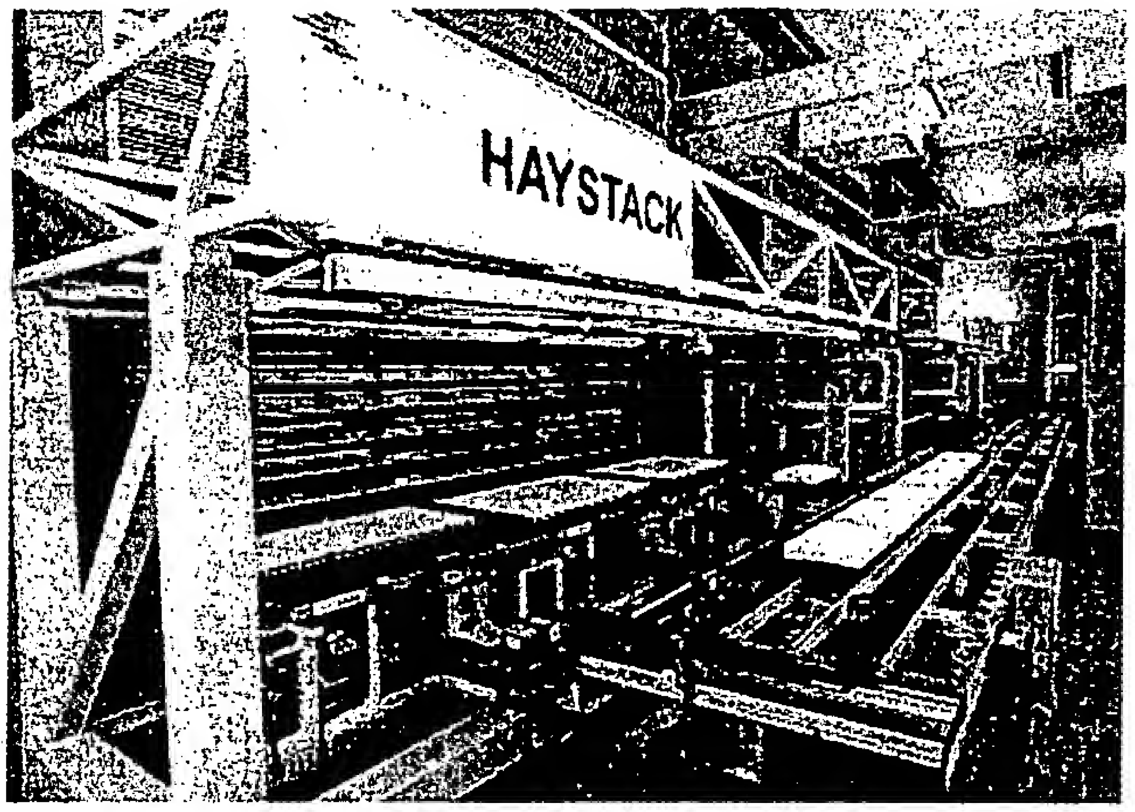


Figure 2. Big is beautiful. The factory-like HTS system maketed by the Automation Partnership.

with the system. This approach is somewhat unusual, as most HTS providers sell equipment outright or provide screening services on a contract basis.

Although the assay systems are getting smaller, the machines handling them still dwarf the benchtop equipment to which most researchers are accustomed. Archer's company markets systems that are "about 150 feet long, about 30 feet high, and weigh 200 tons." Systems on this type of scale are very alien to most researchers working in R&D programs, who are used to working in more of an academic atmosphere rather than a factory-like environment, in which high numbers of repetitive operations are performed around the clock.

Some observers downplay this problem, but Archer contends that it will be a key issue in the development of HTS, suggesting that companies should establish physically separate, factory-like HTS facilities with dedicated staffs[2]. At least one company, Bristol-Myers Squibb (Princeton, NJ), seems to have taken that advice to heart, building a new wing to accommodate an HTS system.

Assay, assay, assay

Whether HTS is carried out in a laboratory or a "factory," building the capacity to perform assays rapidly and reliably may not be the most serious hurdle. Ultimately, it is the assay itself that will determine the success and reliability of screening results, and thus companies are actively searching for new systems to supplement conventional drug activity tests. One area in which several companies are working is the use of screening approaches based on gene expression patterns to find novel drug targets.

According to Richard Shimkets, director of internal research at CuraGen (New Haven, CT), many of the traditional gene profiling approaches (e.g., serial analysis of gene expression, differential display, or subtractive hybridization) are not amenable to high-throughput applications. What's more, it is difficult to gauge the impact of current DNA array-based technologies on HTS: "I think this is an area that people are starting to commit to, but I think they've barely scratched the surface of it," says Robert Lipshutz, vice president of corporate development at Affymetrix (Santa Clara, CA), a supplier of expression profiling DNA chips. Indeed, Mark Benjamin, senior director of business development at Rosetta Inpharmatics (Kirkland, WA), is skeptical about the long-term prospects for standard DNA arrays in HTS. "The first steps require exposing cells and then isolating RNA, which is something that's very hard to do high-throughput," he says.

Another drawback is that most of the useful drug targets are likely to be unknown (particularly in the agricultural sciences where genome sequencing has only just got underway), and DNA arrays that are currently available test only for previously sequenced genes. Indeed, some argue that current DNA arrays may not be sufficiently sensitive to detect the low expression levels of genes encoding targets of particular interest.

To address some of these issues, Shimkets and researchers at CuraGen have developed a rapid restriction enzyme-based mRNA profiling technique (*Nat. Biotechnol.* 17, 798–803, 1999), in which cDNA prepared from mRNA is digested with individual pairs of restriction enzymes, amplified with PCR, and the lengths of the resultant fragments compared with known sequence fragments lodged in a database. Importantly, the presence of unpredicted fragments flags novel genes, thus allowing differentially expressed genes to be cloned, even if they are not represented in expressed sequence tag (EST) libraries.

Researchers at Aurora Biosciences are using genome-wide, gene-tagging strategies to identify novel genes and study expression of known genes in cell lines in response to drug treatment. These assays employ a promoter-trapping strategy using a β-lactamase reporter and fluorescent substrate analogs to measure promoter activity. The resultant β-lactamase-tagged cell library is subjected to fluorescent-activated cell sorting to rapidly separate clones where the tagged gene is constitutively expressed, and sequential rounds of sorting can be used to identify cell clones whose genes are induced or repressed by different drug candidates.

Whereas Aurora's approach primarily makes use of mammalian cells, other companies are pursuing yeast genetics to develop high-throughput assays. Variants on the yeast two-hybrid system, which is widely used for analyzing protein–protein interac-

Cutting screening down to size

One way to do more screens in less time is to pack more on a plate. By carrying out assays in smaller volumes, significant savings can be achieved in the cost of targets, compounds, and reagents. The miniaturization of an assay screen also allows more assays to be carried out in parallel in the same space. Lev Leytes of LJL Biosystems (Sunnyvale, CA) puts it succinctly: "Combinatorial compounds come in small quantities, and targets are also expensive and take time to resynthesize. An assay carried out in 3 ml rather than 100 ml provides a 97% saving, which can amount to a significant sum of money." Some of those savings are already being transferred to drug companies. According to Douglas Kell[1], a typical screening campaign has dropped in cost from US$35 million to US$1.1 million. And assays are continuing to shrink. The new 1,536-well microplate already has competition from laboratory-on-a-chip systems, which use electrokinetic forces to move and mix cells/reagents through etched microfluidic channels/reactor vessels as small in diameter as 10 µm. Both Orchid Biocomputer (Princeton, NJ) and Caliper Technologies (Mountain View, CA) have developed such chips, using different proprietary technologies. But the move toward miniaturization also poses its own problems. Reagents such as DNA behave like concrete in solution, making them difficult to move around; evaporation becomes significant in microliter volumes, and capillary action causes problems with "wicking" and bridging of liquid between wells. What's more, although miniaturization allows more precise control of environmental fluctuations, the relative impact of fluctuations becomes huge. Statistics will also place an additional limit on miniaturization. "Our current reactors contain about 1,000 to 2,000 cells [per well], so we can get down probably at least an order of magnitude [smaller]," says Sheila Dewitt, senior director of business development at Orchid. Beyond that point, the wells may not contain enough cells to produce statistically significant results.

Alan Dove and Andrew Marshall

Measure for measure

Most HTS assay readouts are currently based on variants of fluorescence measurement. These include fluorescence polarization, time-resolved fluorescence, fluorescence resonance energy transfer (FRET), and fluorescence correlation spectroscopy (FCS). With the move toward miniaturization (see "Cutting screening down to size"), increases in assay sensitivity will require new technologies (as assay volume is reduced, sample signal decreases while background signal increases). According to Lev Leytes, president and CEO, LJL Biosystems is developing a variant of time-resolved fluorescence that "can find sample signals that are as little as 5% of the background." FCS, another relatively new analytical technology, also allows binding properties to be determined at the level of single molecules and in volumes as low as a femtoliters, all in a matter of milliseconds. Other groups are working on ways of integrating sensors into miniaturized assay systems that can increase sensitivity and reproducibility. For example, David Walt and his colleagues at Tufts University (Medford, MA) have designed fiber-optic sensors (*Nat. Biotechnol.* **14**, 1681–1684; 1996) that can be arrayed and etched into the bottom of microwells, allowing assays to be performed in extremely small (femtoliters) volumes. As most fluorimetric assay systems currently employ a single detector to interrogate wells on a microplate sequentially, throughput of the reader is another issue. New work suggests that coupling a telecentric lens and CCD camera can save substantial time by allowing entire plates to be imaged and read simultaneously[1]. Apart from fluorescence, several other nonisotopic assay methodologies are currently being assessed for HTS, including surface plasmon resonance (which provides information on all sorts of different types of binding measurements in real time), infrared spectroscopy (which can provide noninvasive measurements of cell changes without the need for exogenous reagents), and more traditional colorimetric and amperometric methods.

Andrew Marshall

tions in vivo, are now under development at several pharmaceutical companies to screen for small molecules[3,4]. Many of these systems attempt to overcome some of the traditional problems of two-hybrid systems that make them unsuitable for HTS, such as false positives and complex assay preparation[3].

Other types of functional screens in yeast are also under development. For example, Cadus Pharmaceuticals (Tarrytown, NY) is functionally inserting mammalian G-protein-coupled receptors (GPCRs) and associated signaling proteins into the pheromone response pathway of yeast to develop screens for identifying modulators that act on components of the GPCR pathway (see p. 878 for details).

To approximate in vivo systems more closely, screens based on the nematode *Caenorhabditis elegans* have also been developed. The small size (it consists of 98 cells and is only 1 mm in length), easy culture, and extensive knowledge regarding *C. elegans* biology makes the worm readily amenable for use in HTS. Companies such as Axys Pharmaceuticals (S. San Francisco, CA) and Exelixis Pharmaceuticals (S. San Francisco, CA) currently use automated nematode screens to evaluate the organism's anatomy and behavior, both before and after genetic manipulation, in response to drugs.

Elsewhere, researchers are designing more exotic assays to supplement cellular assays. For example, Chad Mirkin of Northwestern University recently reported a system that exploits the optical properties of nanometer-sized gold microspheres. Mirkin claims that microspheres derivatized with nucleic acids show reproducible and robust chromophoric changes on ligand binding[1]. Other laboratories are also developing assays that allow lipid-soluble proteins (e.g., G-protein receptors that are intractable to conventional assays) to be screened by array on isolated membranes.

Another challenge is to design reliable assays to monitor a target gene product's expression or activity. In this context, some of the technologies lumped under the term "proteomics" may prove useful. Microarray and chip-based protein binding assays like those being developed by Ciphergen (Palo Alto, CA) and Large Scale Biology (Rockville, MD) may facilitate high-throughput target identification, whereas the mass spectrometry approaches being pursued by companies like Oxford Glycosciences (Cambridge, UK) might provide more detailed information on individual targets. The preliminary nature of most current technology in proteomics, however, makes it difficult to predict the impact of this field on HTS (*Nat. Biotechnol.* 17, 233–236, 1999).

Lastly, a useful assay must be compatible with a company's existing automated HTS system. In an effort to integrate assay development and execution, some manufacturers of HTS equipment produce benchtop versions of their systems, but the rapid development of the field and the number of vendors producing equipment make it difficult to determine whether assay scaleability will become a serious stumbling block in the future.

Do It Yourself HTS

Many pharmaceutical firms are relying on smaller, specialized companies to provide equipment and even whole turnkey systems (see Table 1 and text), but there may be advantages to keeping HTS development in-house. Buying off-the-shelf components and cobbling them into an integrated screening system carries its own set of risks, but the ability to retain control over databases and proprietary compound libraries has strong appeal. Glaxo Wellcome (Research Triangle, NC) is one of the few large pharmaceutical companies that has not entered partnerships with HTS companies. H. Mario Geysen, a research administrator at Glaxo, explains that "you can build up your own robotics. You buy the pieces and you set up a shop and you just link them together so that it fulfills the function that you want to execute." Geysen claims that the difficulty of integrating robotic systems lies primarily in interfacing different computer languages, a problem that the company's programmers have generally been able to overcome. Human Genome Sciences (HGS; Rockville, MD) has taken a different tack, developing its own HTS facility from the ground up. By isolating full-length human cDNAs corresponding to genes that produce putative secretory proteins—akin to the signal sequence trap libraries developed by Genetics Institute (Cambridge, MA) and Genentech (Paris)—HGS has established a protein expression library that can be screened for desired activities. Unlike small-molecule HTS programs, HGS's biological HTS efforts have the advantage of starting with a relatively small, well-defined library of proteins, "so we can afford to do a type of cell biological screening that is very rapid, but in some cases very deep," says company president Bill Haseltine.

Alan Dove

The eve of ADME

If novel targets can be found and new assays developed, the next bottleneck in HTS will be at the other end of the process: testing the new lead compounds to determine their toxicity and metabolism. Activating or inhibiting a particular gene product in cultured cells is a far cry from treating a disease in a whole organism, and most assays for pharmacokinetics and toxicology currently rely on expensive, time-consuming animal tests.

Accelerating studies on the absorption, diffusion, metabolism, excretion (ADME) and toxicology of a drug poses a monumental obstacle, but one that some HTS companies are hoping to surmount.

"Most lead optimization is done the same way primary screening was done 10 years ago—it's benchtop biology," says Lansing Taylor, CEO of Cellomics (Pittsburgh, PA). As automated HTS churns out increasing numbers of lead compounds, ADME researchers are finding themselves overwhelmed. In collaboration with the optics manufacturer Carl Zeiss (Thornwood, NY), Cellomics has developed automated assay systems that provide a detailed view of the behavior of gene products in whole cells under different conditions, an approach that Taylor refers to as "high-content screening [or HCS]."

The idea is to automate the experiments that would be the focus of early-stage ADME and toxicology studies, and to weed out compounds with unsuitable behaviors in whole cells. Rather than replacing existing assays, the Cellomics technology is designed to be incorporated into existing HTS systems. After a "first round" high-throughput screen to identify compounds with a desired activity, flagged compounds enter a round of more complicated tests for toxicity. This second round is referred to as "high-content screening" or HCS because of the amount of additional information it provides. "If you have. . .a 1,536-well plate, and a 0.5% hit statistic, that means you're going to have seven or eight hits per plate. That plate is now transferred into the HCS reader, which now only has to read seven or eight wells," Taylor explains.

DNA microarrays have also been touted as a tool for toxicological studies, since toxicity generally involves changes in gene expression patterns, but the arrays used for general-purpose screening will probably be unsuitable. Spencer Farr, CEO of Phase I Molecular Toxicology (Santa Fe, NM), cautions that "you certainly don't want to walk into the FDA and say, 'With this compound, we saw the following 27 genes turned on 30-fold, but we have no idea what they are!'" Instead, a few HTS companies are trying to develop arrays of surrogate markers that are known to be activated by particular types of toxins.

Affymetrix now produces a dedicated toxicology array based on a rat model, and Phase I is pursuing several approaches to high-throughput toxicology. "We try and put together assays where we glean more than just the old 'when does it die?' kind of information. More importantly, why is it sick?" says Farr. In addition to surrogate marker arrays, Phase I has also developed hypersensitive cell lines that lack crucial repair machinery.

Whether the systems are designed for ADME or toxicological assays, the guiding principle is to "fail fast and fail cheap," eliminating unpromising lead compounds from consideration as quickly as possible; howev-

Table 1. Technologies and customers of some major high-throughput screening companies.

Company	Major HTS products	Major partners/customers
Aurora Biosciences (San Diego, CA)	High-sensitivity fluorescence assays and UHTS on 3,456-well plates	Warner-Lambert, Bristol-Myers Squibb, Eli Lilly, Hoffmann-LaRoche, Pharmacia & Upjohn, Becton-Dickinson, Cytovia
Automation Partnership (Cambridge, UK)	Large-scale, turnkey drug discovery factory, equipped for sample storage, automated cell-culture, screening, and database management	Merck, Bristol-Myers Squibb, SmithKline Beecham, Zeneca, Proctor & Gamble
Axys Advanced Technologies (South San Francisco, CA)	Manufactures combinatorial libraries for HTS, Nematode HTS systems	Daiichi, Parke-Davis, Rhône-Poulenc Rorer, Pharmacia Upjohn, Sigma Pharmaceuticals, Protein Design Labs
Caliper Technologies (Mountain View, CA)	Uses silicon microchip manufacturing technology for miniaturized HTS systems	Hoffmann-LaRoche, Hewlett-Packard, Dow Chemical
Cellomics (Pittsburgh, PA)	High-throughput 3-dimensional cell imaging system to screen primary "hits" for information about cellular activity and toxicity	Johnson & Johnson, Merck, Parke-Davis
CuraGen (New Haven, CT)	Gene expression assay system integrated with databases of information on individual genes	Biogen, COR Therapeutics, Genentech, Glaxo Wellcome, Hoffmann-La Roche, and Pioneer Hi-Bred International
LJL Biosystems (Sunnyvale, CA)	High-sensitivity fluorescence assays and assay readers	SmithKline Beecham, Monsanto, Pharmacia & Upjohn
Orchid Biocomputer (Princeton, NJ)	Uses silicon microchip manufacturing technology for miniaturized HTS systems	Beckman Coulter, Advanced Bioanalytical Services, SmithKline Beecham
Pharmacopeia (Princeton, NJ)	Combinatorial chemistry libraries, molecular modeling software and services for drug discovery	Libraries licensed to Bayer, Novartis, Schering-Plough, Zeneca; drug-discovery collaborations with Akzo Nobel, Bayer, Berlex Laboratories, Bristol-Myers Squibb, Daiichi,Pharmacia & Upjohn, Schering-Plough
Phase I Molecular Toxicology (Santa Fe, NM)	Assays for toxicological analysis of primary HTS "hits"	Partnerships with Clontech and Amersham; clients include Novartis, Hoffmann-LaRoche,Genentech, Hoechst Marion Roussel,Bayer, Rhône-Poulenc Rorer
Scriptgen Pharmaceuticals (Waltham, MA)	Assays for drug target identification and target interaction studies	DuPont, Eli Lilly, Monsanto, Hoechst Marion Roussel, Hoffmann-LaRoche

er, the preliminary nature of the work in this area suggests that lead optimization is a long way from achieving this goal at present.

Deconvoluting the data

Although HTS companies differ with regard to the relative significance of the bottlenecks in ADME, assay development, or hardware integration, there is widespread agreement that the biggest obstacle in HTS is collating, deconvoluting, analyzing, sorting, and storing the information derived from the assays (i.e., bioinformatics).

Although it is referred to as a single problem, bioinformatics is likely to create several distinct obstacles to the development of HTS. Mark Benjamin at Rosetta explains that the first hurdle is the deceptively simple problem of information storage and retrieval: "Everyone who is using more than a couple of microarrays a day is saying 'I can't store all this stuff in Microsoft Excel.'" Microarrays, which may produce in excess of 100,000 data points for a single experiment, are not the only source of trouble. Even "simple" HTS programs are often testing combinatorial libraries of more than a million compounds, easily overwhelming most standard database software.

One solution is for the HTS companies themselves to develop the informatics tools their customers will need. CuraGen, which carries out HTS for hire, analyzes a customer's samples with its proprietary screening technology, then provides the results and database services electronically. "Everyone accesses this information through the same Internet connection," says Shimkets.

The future of many HTS companies may well lie in subcontracting for pharmaceutical companies, but many contend that large companies will want to keep their HTS activities to themselves (see "Do-it-yourself HTS"). For in-house programs, companies like Rosetta Inpharmatics are hoping to provide bioinformatics tools and products that customers can use as they see fit. Rosetta recently acquired Acacia Biosciences (Richmond, CA), adding that company's yeast bioassay to a technology portfolio that already included DNA arrays. Mark Benjamin asserts that Rosetta's bioinformatics products are designed to facilitate the use of its arrays, whether the arrays are used for HTS, early stages of assay development, or basic research. Rosetta also plans to market its software as a stand-alone product for use with other brands of microarrays.

One of the other brands that benefits from separate software is Affymetrix. In contrast to Rosetta's approach, Affymetrix has focused on producing chip hardware, which is sold with software that converts results from the chip into a standard data format called GATC (Genetic Analysis Technology Consortium). Since the standard has been made public, other companies are free to write software to analyze Affymetrix chip output.

Competition and specialization may be necessary to address the next bioinformatics hurdle in HTS. "I'm collecting all of these data, [but] I'm not really interested in storing and retrieving and analyzing. What I [really] want to know is how can I use this to actually design a better drug?" says Benjamin. Affymetrix's Lipshutz agrees, adding that the capabilities of microarray hardware are now being restricted by bioinformatics problems: "It's not the assay per se that's useful, it's basically understanding how to interpret the data biologically."

Growing pains

Ultimately, the test of HTS will be whether the technology can live up to its hype. "The worry we have is that HTS is not only important in the commercial sense, but it's important to be seen working in the area," says Archer. As a result, some pharmaceutical companies may be pouring resources into HTS programs without carefully considering the limits of the approach and the serious obstacles it still faces. "Wall Street is going to be asking not what machines have you put into your facility, but how many hits per day are you generating in reality. If [HTS] fails to deliver. . .it will fall into disrepute for reasons that have nothing to do with the technology," says Archer.

Indeed, a problem that has remained largely unaddressed by current HTS efforts lies outside the technology, in the infrastructure and attitudes of traditional pharmaceutical R&D programs. The model of academic research, in which small groups collaborate on different aspects of a problem while retaining some independence from each other, may be a recipe for disaster when applied to HTS. If an assay development team is not intimately familiar with the limitations of automated screening equipment, new assays may force the screening group to retool a roomful of equipment. Similarly, a library of compounds built around a common backbone that is poorly absorbed in vivo may produce hundreds of hits in a preliminary screen, only to have them eliminated in late-stage ADME tests.

1. Kell, D. Screensavers: trends in high-throughput analysis. *Trends Biotechnol.* **17**, 89–91 (1999).
2. Archer, R. Towards the drug discovery factory. *J. Assoc. Lab. Automat.* 3, 4 (1998).
3. Frederickson, R. Macromolecular matchmaking: advances in two-hybrid and related technologies. *Curr. Opin Biotechnol.* **9**, 90–96 (1998).
4. Young, K. et al. Identification of a calcium channel modulator using a high-throughput yeast two-hybrid screen. *Nat. Biotechnol.* **16**, 946–950 (1998).

JOURNAL OF BIOMOLECULAR SCREENING
Volume 3, Number 4, 1998

Review Article

A Review of Automation Options to Support Plate Preparation, Cherry Picking, and Homogeneous Assays

JASON W. ARMSTRONG,[1] RICK A. GERREN,[2] and STEVEN D. HAMILTON[2]

ABSTRACT

Developments in high throughput screening (HTS) have led to new needs in automation to enable better handling of applications such as homogeneous assays and cherry picking. Software and hardware integration approaches for screening automation have been changing in concert with these new application needs. The result of this combination has been the production of robotic systems for drug discovery with improved stability and functionality. This review critically assesses some Zymark, Tecan, and Beckman solutions for current HTS requirements.

INTRODUCTION

In the last four years, robotic high throughput screening (HTS) for drug discovery has moved from being a desire of most companies to a reality.[1,2] During this period, robotics and their control software have evolved from systems which were often unreliable and very time consuming to second generation systems which produce far better lab productivity and flexibility (e.g., Zymark Virtuoso™, Hopkinton, MA; Beckman Core Systems™, Fullerton, CA; Tecan Genesis™, Switzerland). However, there is further room for improvement in primary screening robotic systems. Promise for improvements may come from newly developed ultra-high throughput screening (UHTS) technologies, e.g. Aurora, (San Diego, CA), Pharmacopeia (Princeton, NJ) and such products as the Zymark Allegro™ (Hopkinton, MA) which take a more industrialized "assembly-line" approach to microplate processing.[3,4,5] Reducing screening cost and increasing throughput are the primary drivers for novel HTS technologies and more robust industrialized robotics.

The successful implementation of primary screening in recent years has resulted in new needs for automation, particularly for secondary screening. These needs stem from the nature of secondary screening, which generally requires more elaborate microplate preparation procedures. Additionally, secondary screening is greatly facilitated by automating compound "re-array" based on the discovery of hits in primary screening (also called "cherry picking"). Challenges in areas of plate preparation, daughter plate creation and "cherry picking" include software and hardware flexibility, pipetting accuracy, system speed, and data tracking. Some robotic systems that perform these tasks are also capable of performing homogeneous assays. The purpose of this article is to review the current commercial options for these tasks based on needs, features, and cost. The systems we will cover include the Zymark Virtuoso™ (Hopkinton, MA), Tecan Genesis configurations, Sagian Core Systems™, and Sagian Core Generations™ (Fullerton, CA).

GENERAL SYSTEM OVERVIEW

Zymark Virtuoso™ (Figs. 1 and 2)

This new product from Zymark is a departure from the company's past approach of custom systems. These systems can be configured at the time of purchase based on a selection of available modules/devices. A common configuration which

[1]HTS Consulting Ltd., 2238 Adrian St., Thousand Oaks, CA 91320.
[2]Source Biopharmaceuticals Inc., Boulder, CO 80304.

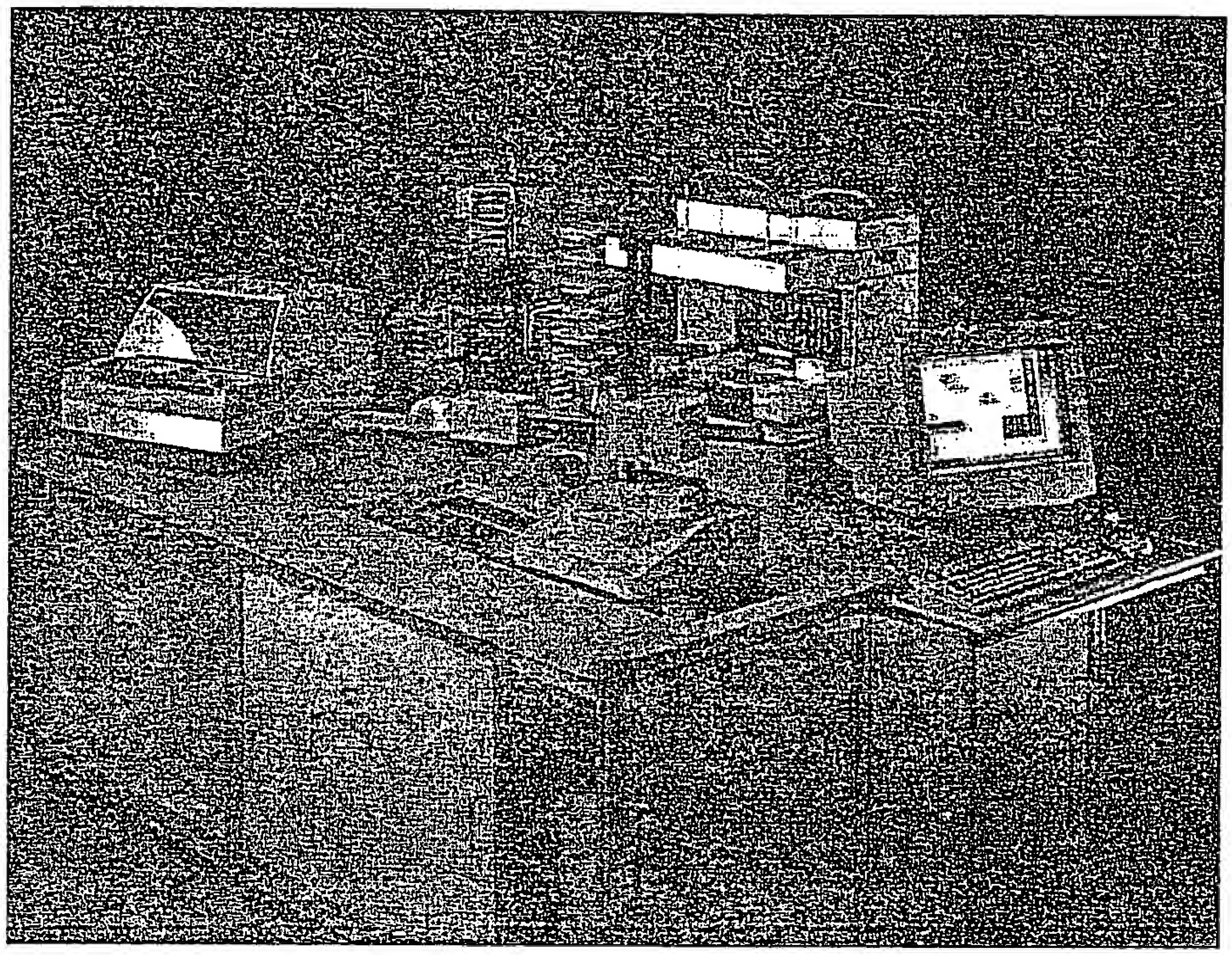


FIG. 1. Zymark Virtuoso 1. This standard product by Zymark is best used for "cherry picking" and homogenous assays tasks. This photographed system does not have the optional Rapidplate 96/384 installed. The normal position of the Rapidplate 96/384 is shown in the diagrammatic representation of Virtuoso 1 (Fig. 2). (*Photograph courtesy of Zymark.*)

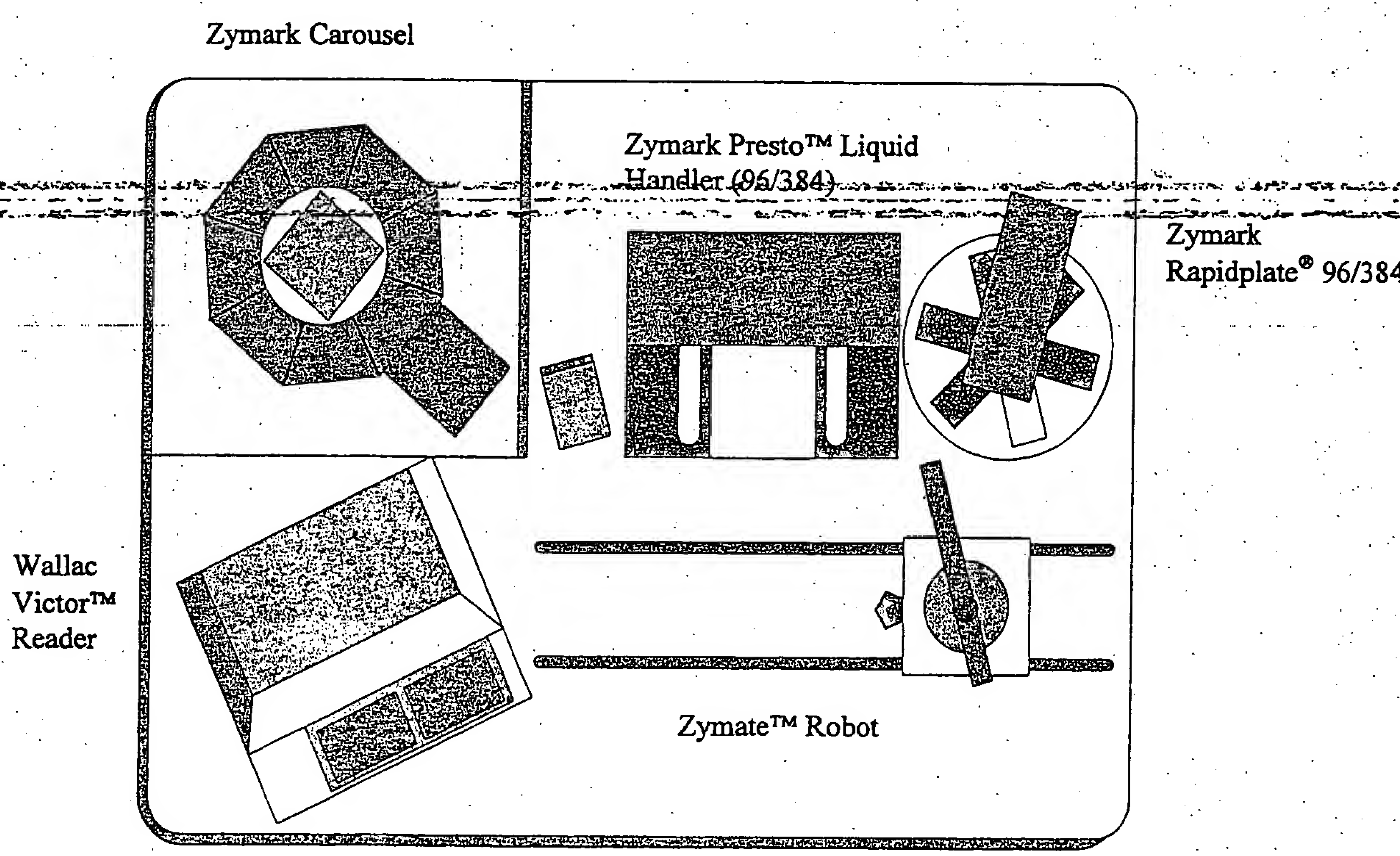


FIG. 2. Diagrammatic representation of the Zymark Virtuoso 1. All possible modules for a Virtuoso 1 configuration are shown.

pharmaceutical companies have begun purchasing for sample preparation, mother-daughter plate creation and "cherry picking" includes the following hardware: a 176-plate storage carousel (temperature control is an option); a *Rapidplate*® 96/384 channel pipetter; a *Presto*™ liquid handler with both an eight and a single channel pipetter (a Cavro [Sunnyvale, CA] OEM product with the option of either Teflon cannulae or removable pipet tips); and an EG&G Wallac (Gaithersburg, MD) Victor™ plate reader. The Rapidplate may be configured to dispose of pipette tips after each use, or wash and reuse these tips. Microplates are transported by a track mounted *Zymate*® XP™ robot. Other configurations are certainly possible based on purchaser's module selection from a list of Zymark standardized choices.

Tecan Genesis™

The Genesis has been put to many uses including combinatorial chemistry, plate preparation, and screening. Some reasons for its success include its pipetting accuracy, choice of deck size ranging from 1 to 2 meters, eight independently addressable X, Y, Z cannulae (Teflon-coated washable cannulae or pipette-tip attaching cannulae), and the *ROMA*™ gripper tool for plate transport. Recently, Tecan has offered the ability to customize the Genesis by interfacing plate processing equipment. These options include a carousel (210 microplates), 96-channel pipetter (to date only the Matrix [Lowell, MA] PlateMate™ and Tomtec [Hamden, CT] QUADRA 96™ pipetters have been integrated), a plate washer, and plate readers. These options make the Genesis capable for plate preparation, "cherry picking" and assays.

Sagian Core Generations™

These small and simple integrated systems come standard as a 1 meter ORCA™ robot servicing a carousel and a Multimek™ (96-well pipetter), together with bar code reading capability. The carousel may be configured with racks for microplates (180), deepwell plates (80), or any combination of microplates or deepwell plates. Beckman (via their Sagian division) can add other modules, such as plate sealers and print/apply bar code labelers, making the system a custom 2 or 3 meter rail system purchase rather than the Core Generations product. The Generations system is driven via an easy to use icon-based *GUI* written specifically for the Generations system. Setup and configuration of the Multimek is accomplished via that devices own GUI. The Multimek retains its front panel interface for manual control. Liquid transfer with the Multimek is pipette-tip based for transfers of 1–200 μl. These tips can be discarded after each use or washed and reused. A fixed, washable tip option is available for transferring volumes of 1–20 μl.

Sagian Core Systems™ *(Fig. 3)*

This type of system differs from those above in both size and complexity. These systems can complete microplate preparation and cherry picking tasks. In addition, functionality extends to the automation of complex assays. Sagian Core Systems include such items as a Biomek™ 2000 pipetter, a Multimek 96-channel pipetter, a plate storage carousel, a CO_2 incubator, readers, print & apply bar-coding, and filtration, as well as other devices. A central, track-based ORCA robot transports microplates to/from the various workstations (tracks can be one to three meters in length). These systems come standard with Sagians' SAMI™ icon-based scheduler, which supports optimized scheduling.

SYSTEM STRENGTHS AND WEAKNESSES

Zymark Virtuoso™

Given the potential functionality of Virtuoso, the system has a low price compared to other competing products. Perhaps the strongest feature of this system is the pipetting combination of the Rapidplate 96/384 and the Presto liquid handler. We have found the Rapidplate 96/384 to be the most accurate and reliable fully automated, robot compatible 96/384 channel pipetter on the market for liquid transfers in the 2–150 μl range (1 μl with a 6% CV is possible in multi-aspirate mode). It is also among the least expensive. The Presto Liquid handler offers the reliability of Cavro pipetting devices with a very small footprint and low cost. Based on our experiences, the combination of these Zymark liquid handlers out performs the Beckman Multimek/Biomek combination in cost, unit size, and pipetting ability. The Virtuoso system has its own table for support and cannot be placed onto a laboratory bench.

The Virtuoso is homogeneous assay capable, which separates it from the other simpler systems in the same cost bracket. The Presto liquid handler possesses the Cartesian positional accuracy to address 384-well plates. Microplate "cherry picking" is done via the Presto single-channel pipetting arm, which can be slow if there are a high number of "hits" to be picked from each plate. The Rapidplate is currently both a 96 well-plate and 384 well-plate pipetter. Therefore, all components on the Virtuoso system are 384-well microplate compatible. Currently, 384-well compatibility is very important as many companies are attempting to transition a substantial percentage of their assays to this format in 1998.

Zymark's current method building and scheduling software for Virtuoso is PCS™ 3.0 running on Windows® 95 or NT™. Although this software lacks some of the power of Beckman's SAMI NT/SILAS™ control software, it is simpler to use and learn. Zymark offers as an option "cherry picking" software, which supports the use of both the Rapidplate 96/384 and Presto™ liquid handler to create "hit" plates or daughter plates. For both PCS and the "cherry picking" software, links to common databases and data storage methods are provided. The new "cherry picking" software is easy to use and very visual in its design.

Tecan Genesis™ *(FACTS)*

Eight independently addressable cannulae give Tecan Genesis™ systems the best "cherry picking" 96-well liquid handling mechanics of all devices on the market. This feature allows up to eight "picks" to be done from a single plate before

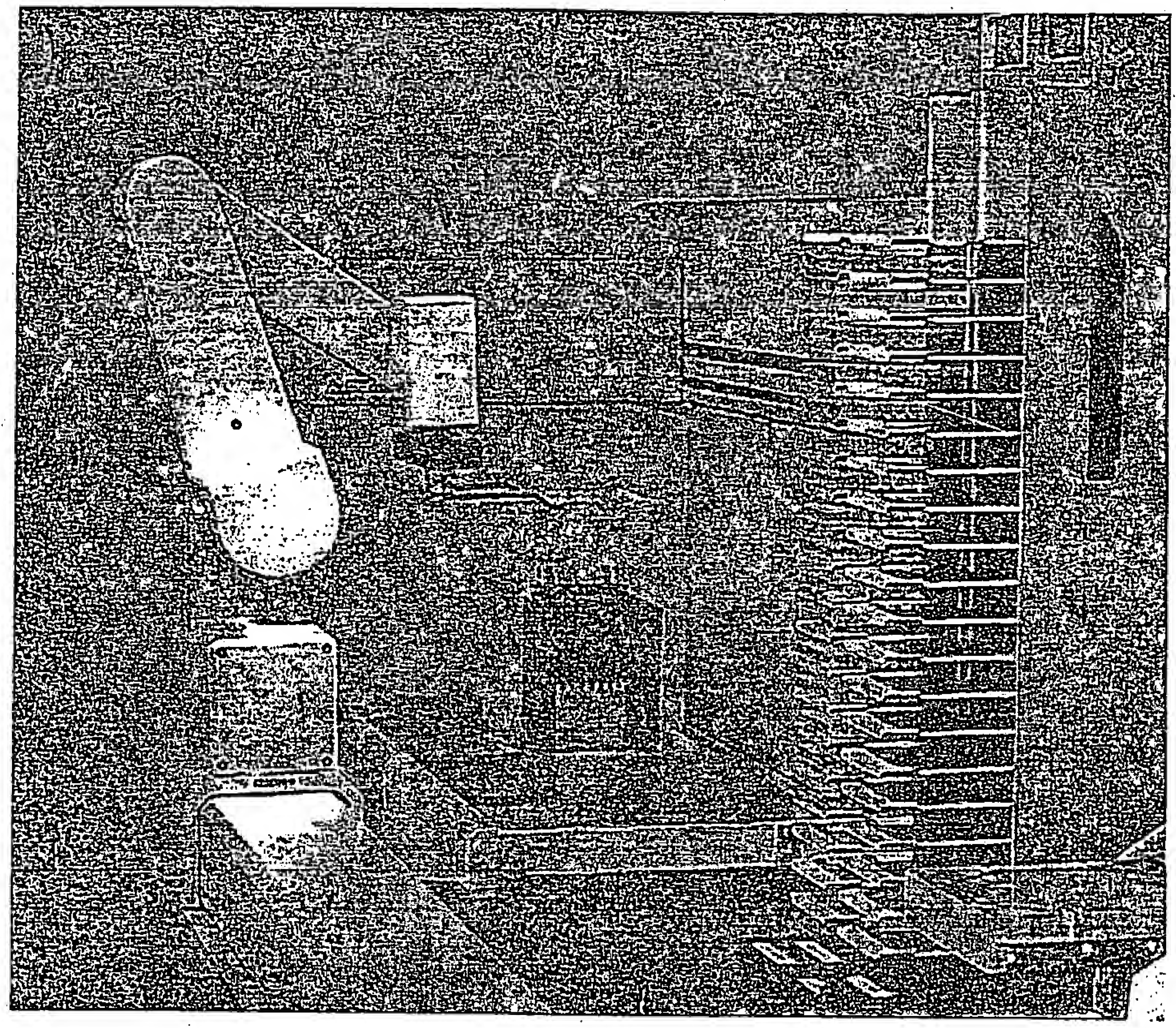


FIG. 3. Beckman-SAGIAN Core System. Visible is the ORCA robotic arm, Biomek 2000 a microplate washer and microplate storage hotel. (*Photograph courtesy of Beckman-Coulter.*)

tip washing is needed. However, maintaining the positional accuracy of eight independent long cannulae may make 384-well addressing a maintenance issue. The Genesis offers the option of both disposable pipette tips or fixed washable cannulae for liquid transfers in the 2–200 μl range. When washable tips are deemed acceptable, this can offer a significant advantage in cost and time savings. The standard Genesis systems lack components which are essential to streamlined "cherry picking" processes. By the very nature of "cherry picking", one has to access many mother plates to create a few daughter plates; therefore, a high-capacity microplate storage device is an important module. Additionally, a 96 channel pipetter is also a useful feature for "cherry picking" systems to enable rapid microplate compound dilutions in a single action. The same components are also very important in an efficient system for plate dilutions and standard mother-daughter plate creation. Once a carousel and a 96-channel pipetter are added, the Genesis FACTS has a high cost relative to the Zymark Virtuoso. At the time of this review, the addition of both a carousel and a 96-channel pipetter is not a standard product for Tecan. The company's experience to date covers only the Matrix PlateMate and Tomtec QUADRA 96 channel pipetters, which could potentially be coupled to a 210 microplate carousel.

The Tecan Genesis can be configured for screening work with combinations of a plate reader (Tecan Spectra™ series) and washer (Tecan 96 PW) and the modules mentioned above. While this is feasible, Tecan has not yet coupled that number of larger modules to one Genesis. Tecan provides software which has good functionality (method builder and a dynamic scheduler), although it is somewhat disjointed (e.g., the pipetting method builder and scheduler do not exist under the same shell, or ".exe" program). As with Zymark's Virtuoso, Tecan "cherry picking" software, with data links can also be purchased. The standard Genesis presents a very tight footprint that easily fits onto a standard lab benchtop. However the addition of a Carousel and 96-channel pipetting device make the overall combination rather large and cumbersome.

Sagian Core Generations™

These systems are generally small and therefore low in complexity and price. Consequently, functionality is limited. This product has been used within companies for mother-daughter plate creation. Such a system configuration includes a carousel, bar code reader and Multimek 96-channel pipetter. Utilizing the ORCA arm, this system occupies a very efficient footprint that could reside on a 36″ lab benchtop. The standard systems are not adaptable to "cherry picking" which would require a Biomek 2000™ (generally only placed on larger Sagian Core systems). Additionally, Core Generations systems are not assay capable, in that an interface to a plate reader or reagent addition system is not part of the Generations package.

In general, we have found that the Multimek is slower at 96-channel pipetting operations than the Rapidplate 96/384. This difference is marked when not reusing/washing pipette tips, as the Multimek tip attaching actions are slow. Exact comparisons

of pipetting accuracy depend on the range of transfer involved and the exact nature of the liquid. Generally, we have had fewer problems with transfer inaccuracy with the Zymark Rapidplate 96/384.

Because of the nature of the Core Generations system software, changes in the hardware configuration are not easily accomplished. For example, the current GUI does not allow the user to compensate for changing racks in the storage carousel between microplate and deepwell types. Instead these changes have to be made in a carousel configuration file. This type of file manipulation can easily lead to errors in method execution, so extra care must be taken when making and saving changes. Beckman/Sagian will customize their standard GUI interface to meet specific customer needs. While this custom interface does not include a scheduler (Sagian provides their icon-based SAMI scheduler for their more complex Core systems), this is not a limitation due to the focused task nature of these systems.

Sagian Core Systems™

These systems provide a wide range of potential functionality and features which are best suited for the automation of more complex procedures, such as heterogeneous screening assays. The SAMI/NT/SILAS™ software architecture offers a highly flexible, powerful, but costly package for system scheduling, control and interfacing. The ORCA and Biomek 2000 mechanics allow space efficient layouts. However, we have found that smaller and simpler automated systems are more suitable for focused, specific tasks. Therefore, for support of secondary screening, for tasks such as plate preparation, and for "cherry picking", the complexity and cost of the larger Beckman systems is difficult to justify with simpler lower cost solutions available.

SUMMARY

The selection of the correct automation equipment is crucial for painless implementation into a drug discovery program. This selection is both a function of vendor options and the nature of the task to be automated. Additionally, not every screening related activity should be automated. Recently, vendors have switched to simpler approaches for automation solutions, stressing standardization rather than customization. The recent convergence of homogeneous assays, robotics designed for homogeneous protocols, and logical plate preparation paradigms will hopefully make the coming years of screening robotics a far more productive investment than previous years. Judging an automation supplier by performance in a previous project of a specific nature, may not be the best strategy for selection of a new technology for a different application. Each of the vendors (e.g., Zymark, Beckman-Sagian, Tecan etc.) have particular screening automation niches in which they are strong. For example, this manuscript describes Zymark as being particularly strong for homogeneous and cherry picking applications. In contrast, Beckman can be viewed as stronger in the arena of heterogeneous cell based assays. Furthermore, technology selection for the same task will vary among purchasers based on differing priorities in throughput, reliability, and cost.

REFERENCES

1. Hamilton, S.D., Armstrong, J.W., Gerren, R.A., Janssen, A.M., Peterson, J.V. and Stanton, R.A. (1996). An overview of automated biotechnology screening. *Lab. Robot. Auto.* 8:287–294.
2. M. Banks, A. Binnie & S. Fogarty. (1997). High throughput screening using fully integrated robotic screening. *J. Biomol. Screen.* 2(3): 133–135.
3. Schumate, C., Beckey, S., Coassin, P. and Stylli, H. (1997). Ultra-High Throughput Screening. *LAN* 2(4):24–29.
4. Rose, D. and Lenmo, T. (1997). Challenges in implementing high-density formats for high throughput screening. *LAN* 2(4):12–18.
5. Alderman, E. and Elands, J. (1998). A novel approach to ultra-high throughput screening. *Gen. Engin. News* 18(2):14–15.

Address reprint requests to:
Jason Armstrong, Ph.D.
HTD Consulting Ltd.
2238 Adrian St.
Thousand Oaks, CA 91320

E-mail: HTS_Jason@ibm.net

2 Robotic equipment and microsystem technology in biological research

Holger Eickhoff, Igor Ivanov, Markus Kietzmann, Elmar Maier, Markus Kalkum, David Bancroft and Hans Lehrach

2.1 Introduction

Each cell of a living organism contains the whole genetic information in form of DNA molecules. The size of the DNA from a single cell, or genome, of human beings is 3×10^9 nucleotide base pairs. Although the DNA information is usually identical in each cell there are several hundred different cell types. This is due to the fact that genetic information is read out from genes and transcribed from DNA into a cell-specific population of mRNA molecules, which itself can be further translated into different types of proteins. Every step of these cellular processes includes complex interactions of DNA, RNA and protein (Alberts, Bray et al., 1994). To understand these interaction mechanisms scientists started to decode the genetic information (Dulbecco, 1986). This task became finally a major goal of the Human Genome Project (Cantor, 1990).

The benefits of this project are visible already before one-tenth of the genome sequence has been revealed. Genomic databases have enabled scientist to access, retrieve and process biological information (Zehetner and Lehrach, 1994). At the same time, the Human Genome Project has changed the attitude and direction of biological research (Tilghman, 1996). Currently, the interest of researchers is focused on finding genes, analysing their expression patterns and their *in vivo* functions as well as further features of the corresponding proteins. The order and the expression profile of biological information is another level of complexity even more important for the understanding of organisms. Genes whose expression is highly specific to a tissue, organ, cell type or disease may be attractive as targets for the development of highly specific therapeutics and diagnostic (Maier, Meier-Ewert et al., 1997).

Since there are approximately 100000 genes predicted for the human genome, new methods and reliable techniques for processing many samples in parallel and at high throughput are needed.

Here we describe how automated robotic systems can facilitate biological research. Robots have been developed mainly for the parallel analysis and the characterization of large DNA array (Meier-Ewert, Maier et al., 1993, 1994). These automated techniques allow the examination of tens of thousands of clones in parallel by hybridisation-based approaches. We also show how to implement the principles of these robotic systems for other biological tasks including protein analysis and the characterisation of gene expression.

2.2 Hybridisation based approaches to genome analysis

Most of the methods for DNA characterisation are based on the fundamental fact that DNA is able to form a full or a partially complementary double helix hybrid from two separate single stranded nucleotide chains (Watson and Crick, 1953). Hybridisation is the interaction of two DNA strands. To detect hybridisation events one strand (target) is usually immobilised on a solid support, e.g. nylon membranes, whereas its counterpart (probe) is fished out by the target from a hybridisation solution. The probe is labelled and the hybridisation is detected by measuring the signal on the solid surface in the region of immobilised target (Wetmur, 1991; Meinkoth and Wahl, 1984).

Hybridisation approaches are important tools for large scale DNA characterisation and require, among other things, upstream clone picking and spotting, the probe hybridisation itself, and downstream image and computer analysis (Lehrach, Bancroft et al., 1997). First of all, a pool of DNA molecules to be analysed is prepared for insertion into bacteria such as *E. coli*. Randomly spread colonies of bacteria are grown on agar plates. Each colony carries a unique DNA fragment or clone. Since bacteria can carry only a relatively short DNA fragment, a large number of clones, or a clone library, is needed for a full coverage of a genome or even a tissue-specific cDNA library. A typical size of a cDNA library is a hundred thousand clones. After picking, selected clones can be grown and kept in microtitre plates. This allows long-term storage, analysis and subsequent retrieval of individual clones. Clones from microtitre plates can be used for DNA amplification by PCR or they can

be arrayed on nylon membranes for subsequent hybridisation with specific probes (Meier-Ewert, Maier et al., 1993).

Automated clone picking

The pattern of colonies randomly grown on agar plates is checked by an image analysis system to address the position of the colonies for picking. The randomly grouped, proportioned and shaped bacterial colonies are automatically selected on the basis of given criteria: colour, shape, size. The image analysis software is able to recognise clones as small as 0.5 mm in diameter and select for blue/white genetic systems in *E. coli*. After defining the colonies' position, software translates coordinates into robot movement for picking. One pin of the picking head (Fig. 1) touches the colony. Then the 96-pins of the picking head transfer and inoculate colonies in a microtitre plate for growth and storage.

We have integrated the picking feature in a flat-bed robot being capable of picking and spotting. In the past 8 years we designed and tested several

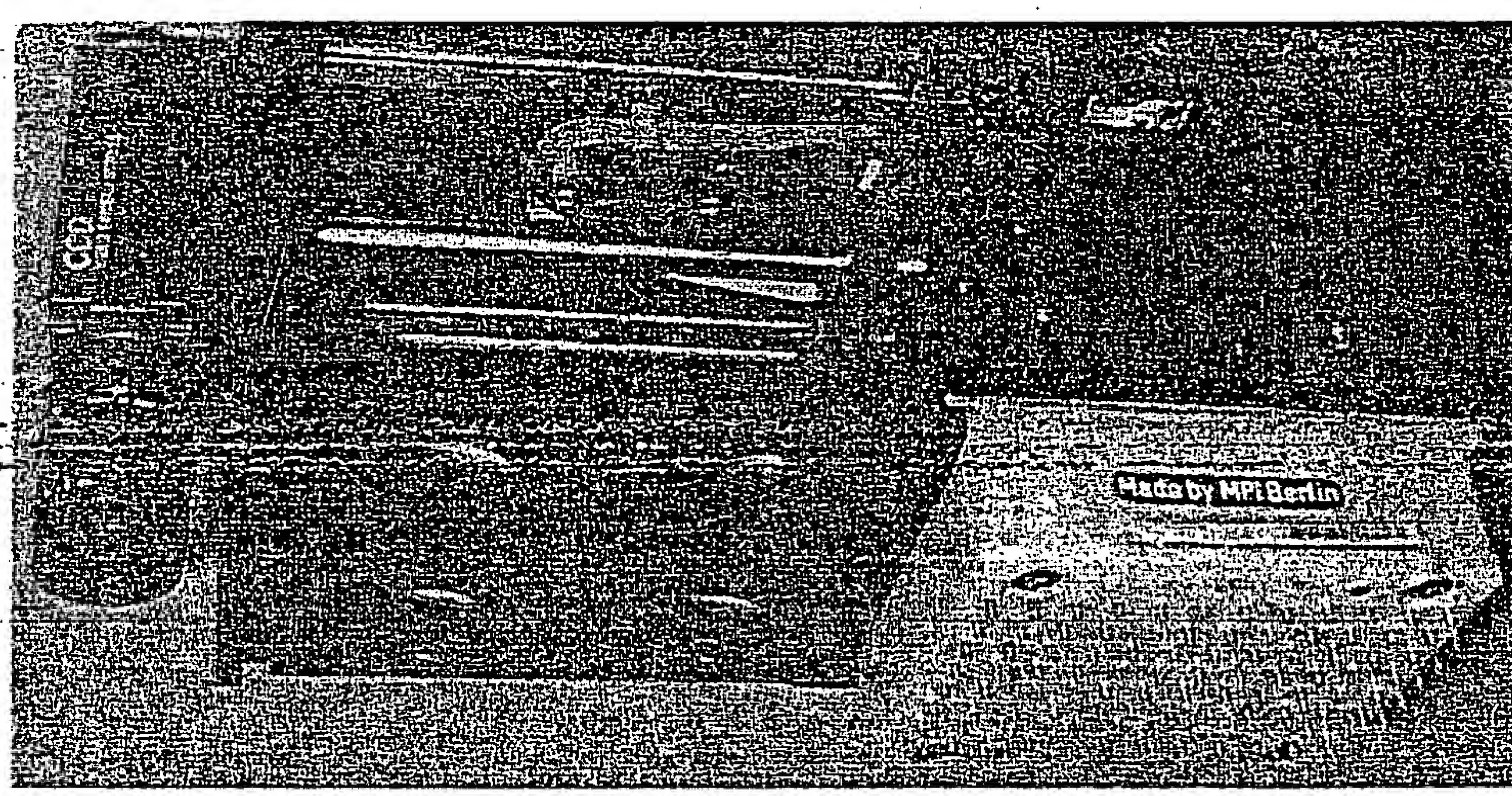


Figure 1 *Picking Head. The picking head consists of three major parts: a CCD camera for taking a picture of an agar colony tray (left side), an x, y moving table that guides the pressure line to the chosen pins and the picking tool with 96 pins*

generations of clone picking systems. The system is capable to pick and inoculate approximately 3000 clones per hour into 384-well microtitre plates; (Maier, 1995).

Robotics systems for automated clone arraying

After picking and growing thousands of colonies in microtitre plates, the colonies are arrayed with a 384-pin head (Fig. 2) onto nylon membranes. The spotting head is moved on a servo-controlled, three-axis, linear drive system with an accuracy of 25 µm. A complete spotting run includes the handling of up to 72 microtitre plates, bar-code reading, lid lifting, 384 parallel clone transfers, pin sterilisation and pin drying. The volume of liquid transferred with a pin depends on the tip diameter, which varies from 150 µm to 450 µm which corresponds from 5 nl to 50 nl liquid volume. The smaller the pin diameter, the higher the spotting density that can be achieved. For routine ope-

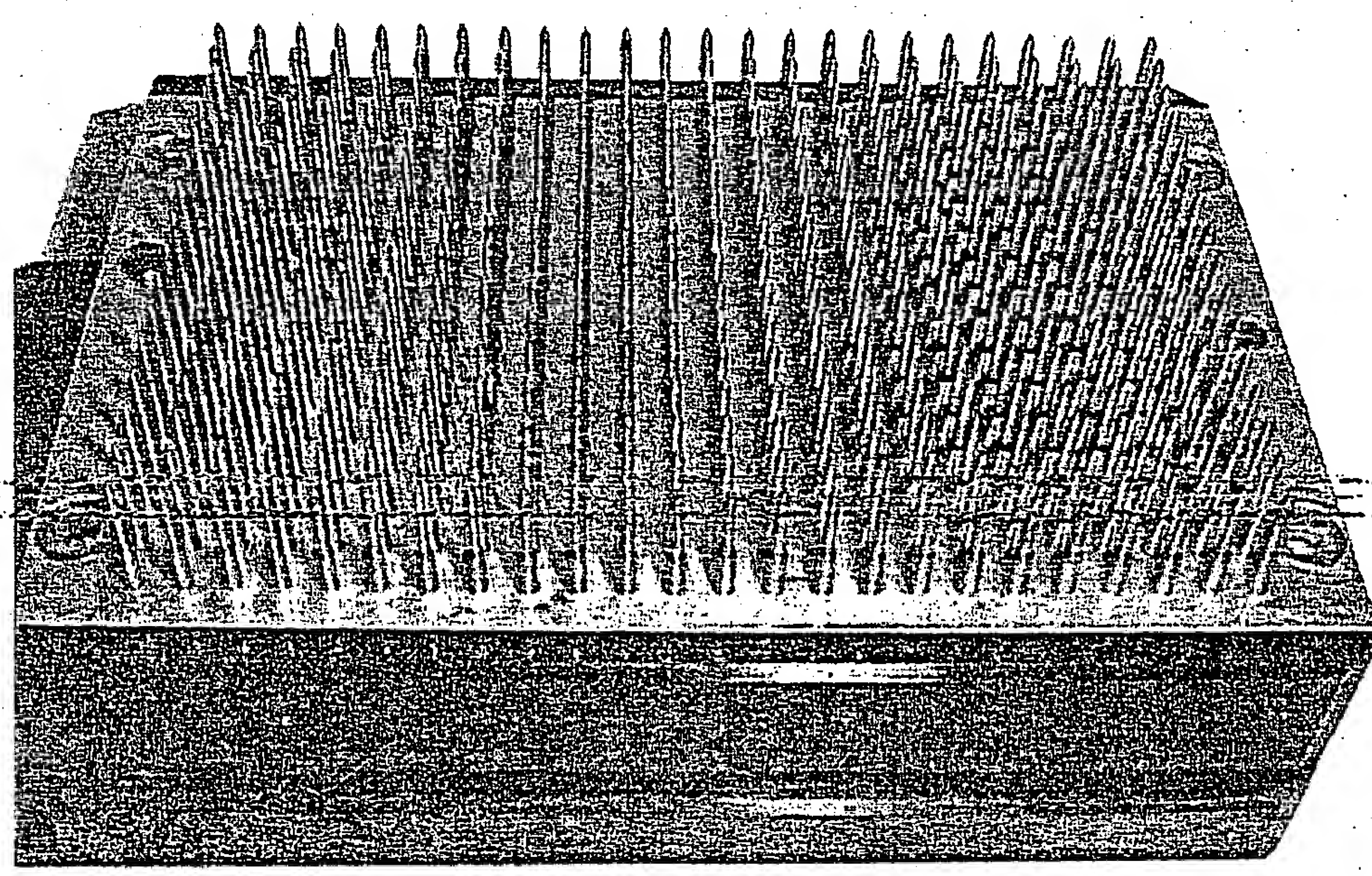
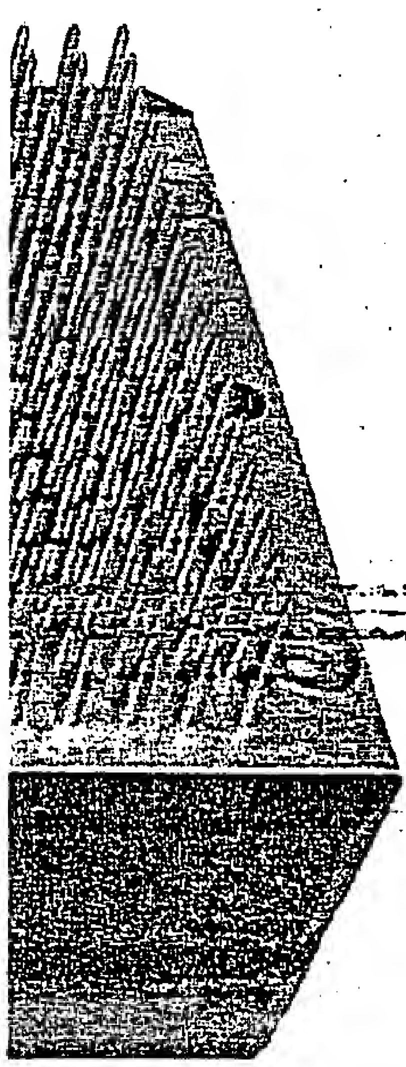


Figure 2 *Spotting Tool. The 384 individually spring-loaded pins are mounted with 4.5 mm spacing in the area of a microtitreplate. The stainless steel pin-tip size varies between 150 µm and 450 µm*

ɔable to pick and
l microtitre plates,

re plates, the colo-
. membranes. The
inear drive system
.es the handling of
384 parallel clone
liquid transferred
150 µm to 450 µm
naller the pin dia-
.. For routine ope-



h 4.5 mm spacing in the
450 µm

rations 27648 clones are spotted in a duplicate pattern as a 5 × 5 box format around 2304 guide dots per 22 cm × 22 cm nylon membrane. The duplicate spotting simplifies detection and identification of positive clones after each hybridisation. The nylon membranes have been reused at least 20 times without significant loss of hybridisation information. With the system described here it is possible to immobilise and analyse up to 147456 clones on a single 22 cm × 22 cm nylon surface.

Large scale thermocycling

In addition to growing arrayed colonies on membranes, suitable DNA amounts can be generated by DNA amplification from colonies, e.g. by the polymerase chain reaction (PCR). PCR techniques (Saiki, Scharf et al., 1985; Saiki, Bugawan et al., 1986) have been developed for a long time and they play a central role in large scale genome analysis programs. Commercially available cycling devices can handle up to four times 384 probes in parallel. We have built a laboratory thermocycling prototype for high throughput DNA amplification based on large water baths (Maier, 1995). A basket filled with 135 different 384-well microtitre plates (51840 reactions) is moved with a pneumatically driven x/z sliding stage between three 220-liter water baths at three eligible temperatures. The microtitre plates are heat sealed with a plastic foil in a commercially available heat sealer to prevent cross contamination. After the amplification step the DNA product is sufficiently pure for spotting (Meier-Ewert, Maier et al., 1993).

Non radioactive hybridisation and detection

The high throughput experiments based on hybridisation require ideally non-radioactive detection methods. At present only a few articles have been published that describe the use of directly labelled fluorescent probes for hybridisation with DNA on solid supports. This is due to the low signal to noise ratio that can be obtained with directly labelled probes. We use a hybridisation protocol which utilises an enzymatic signal amplification (Maier, Crollius et al., 1994). The DNA probe has a tag, e.g. digoxigenin, which is

recognised by an antibody conjugated with alkaline phosphatase (anti-digoxigenin-AP Fab fragment, Boehringer-Mannheim). The hybridisation is visualised with a non-fluorescent substrate for alkaline phosphatase called Attophos. This substrate becomes highly fluorescent after a phosphate group in the Attophos molecule is removed. The detection is quite sensitive since each active centre of alkaline phosphatase can process about 10^4 to 10^5 substrate molecules per minute (Cherry, Young et al., 1994). This labelling method is reliable for a wide range of hybridisation probes, ranging from short oligonucleotides to long PCR products. For documentation and analysis of hybridisation the positive signals are detected by excitation by UV light (365 nm) and photographed with a high resolution CCD camera (Photometrics PXL, KAF 1400 chip) through a fluorescence emission filter (589 nm bandpass filter, bandwidth 80 nm, Herolab, Germany). The pictures are digitised into a Macintosh PowerPC 8100. The obtained spatial resolution is drastically increased in other detection systems that use Time Delay Integration Linescan cameras or laser scanning principles.

Automated image analysis

An important feature in high throughput laboratories is the automated, large scale characterisation of positive clones in the investigated libraries. The main requirements to such an automated analysis system are firstly the automatic grid finding on an array and secondly the determination of positive clones. Unfortunately, different hybridisation arrays show different qualities. The quality of the picture can be affected by several factors, e.g. uneven distribution of spots on a picture due to non flatness of nylon membranes or a high hybridisation background. Human judgement is so far the best method for the decision of whether a clone is positive or not. Nevertheless, the algorithms for an automatic spot finding are quite well developed (Geman and Geman, 1984; Lehrach et al., 1997). At the current stage nearly 80 % of all positive clones can be scored automatically.

2.3 New technologies in high throughput screening: Miniaturisation is a driving force

Integrated circuits made personal computers possible that have revolutionised the world. In addition, the semiconductor industry has been able to double the complexity of a chip every 5 years with reducing cost.

We now witness the development of modern chip-biology, which adopts methods and technologies from the semiconductor industry. High density arrays with integrated solid-phase oligonucleotide synthesis for rapid multiplex analysis of nucleic acid samples have been introduced (Chee et al., 1996; Hacia et al., 1996; Schena et al., 1996). Microlithographic etching of silicon wafers enables the creation of precisely controlled structures, for example "obstacle courses", which might be a good substitution of common gels for separation of long polymer molecules (Volkmuth et al., 1995). Over the last few years, miniaturisation became a driving force in molecular biology and genome analysis.

High throughput screening methods of clone libraries would benefit from further automation and miniaturisation as well. For example, increasing the density of arrays and therefore the number of DNA targets would increase the analysis speed and the information flow. As an alternative to conventional arraying with pins, a microdrop spotting on demand of technology was developed. With the aim of reducing the size of hybridisation arrays by one or two orders of magnitude, the genetic samples are pipetted with a piezoelectric multi-channel microdispensing robot. The piezoelectric dispensing system was originally developed for use in ink-jet printers. The major part is a piezoelectric element or piezoelectronic actor in tube shape, which expands and contracts in the process of the applied AC voltage. This actor is connected to a tapered glass capillary with an outlet nozzle size of 25 µm to 50 µm. The liquid that has to be dispensed can be filled into the glass capillary in two ways. The first possibility is to aspirate the liquid through the dispensing nozzle by applying a gentle vacuum. The second one is to fill the capillary from a reservoir in the back. Once a glass capillary is filled with several microliters, liquid droplets can be shot out by applying alternating voltage. The piezoactor expands and then contracts so that a droplet is fired out of the glass capillary (Fig. 3). The microdrop system is able to dispense 25 µm to 100 µm droplets

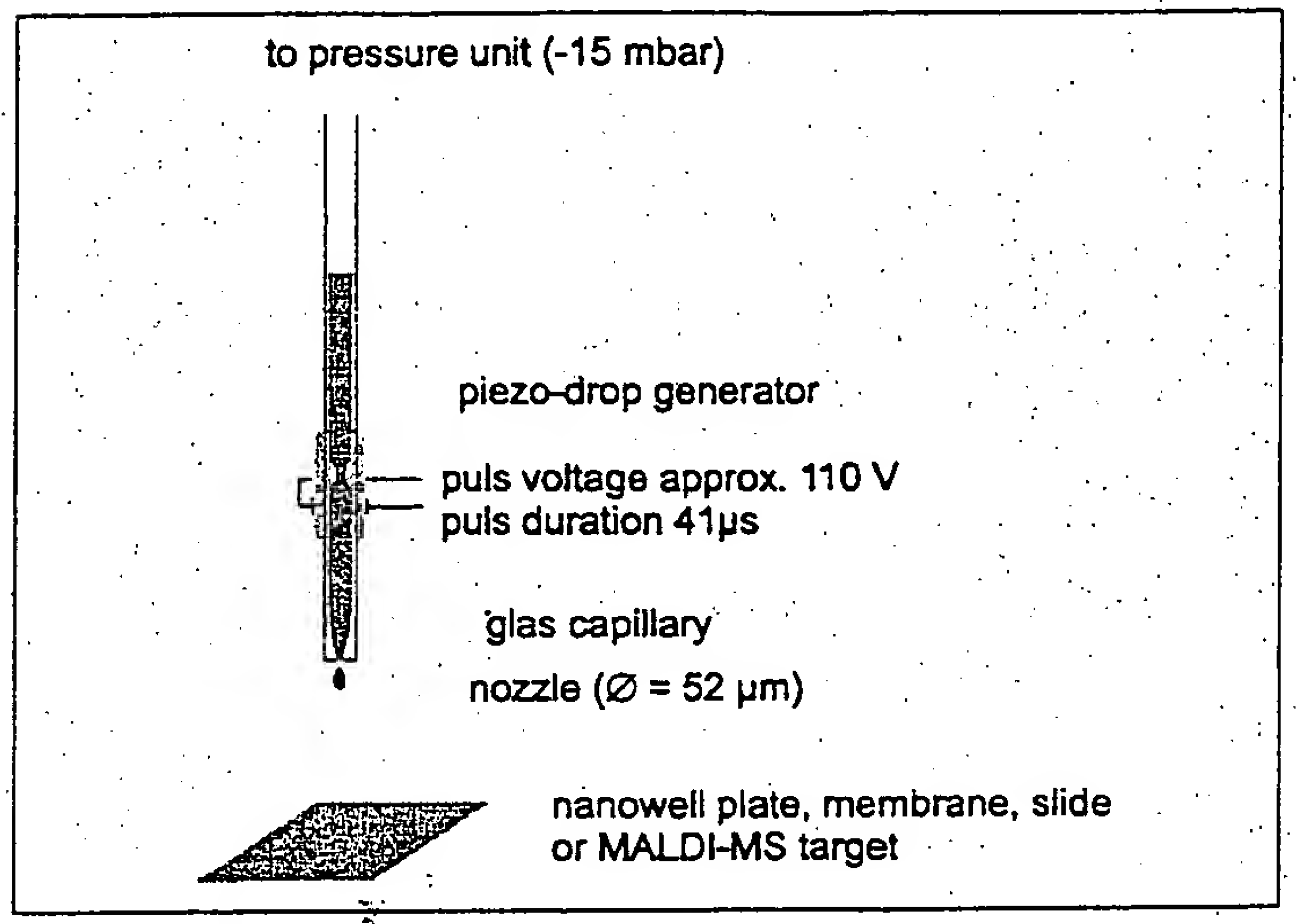


Figure 3 *Scheme of a Piezo-Jet dispenser. A minimum of 11 µl of liquid has to be aspirated. Application of the right voltage together with the right pulse shape to the piezoelectric element results in a drop with 60 µm diameter, corresponding to 100 pl. The drop velocity is approx. 0.58 m/s*

(10 pl to 100 pl). The frequency of a commercially available single nozzle system is approximately 2000 drops per second. The droplet shape is influenced by the diameter of the capillary and the cleanliness of the nozzle edge. Small crystals or a soiled tip surface result in poorly shaped droplets or no droplets at all.

An eight-nozzle head has been recently implemented. It moves in the x, y, z positions with 5 µm resolution using a servo-controlled, linear drive system. A 9-mm spacing between nozzles enables aspiration of solutions directly from the wells of a 384-well or 96-well microtitre plate. After aspirating the samples, dispensing nozzles move to a camera to check whether a suitable droplet is formed or not. The camera captures an image of a droplet in a stroboscopic light. Integrated image analysis system scans the image to verify the quality of the droplets. If a droplet is poorly formed, the image analysis system directs the head to clean the edge of the nozzle. The piezo-dispensing parameters, e.g. the voltage and impulse length for each of the nozzles, are independently controlled (Fig. 4).

With the system described here it is also possible to perform precise filling procedures for nanotiter plates or silicon wafers. Therefore, a second video

able single nozzle
t shape is influenc-
f the nozzle edge.
ed droplets or no

t moves in the x, y,
inear drive system.
tions directly from
spirating the sam-
t a suitable droplet
t in a stroboscopic
erify the quality of
ysis system directs
nsing parameters,
are independently

orm precise filling
re, a second video

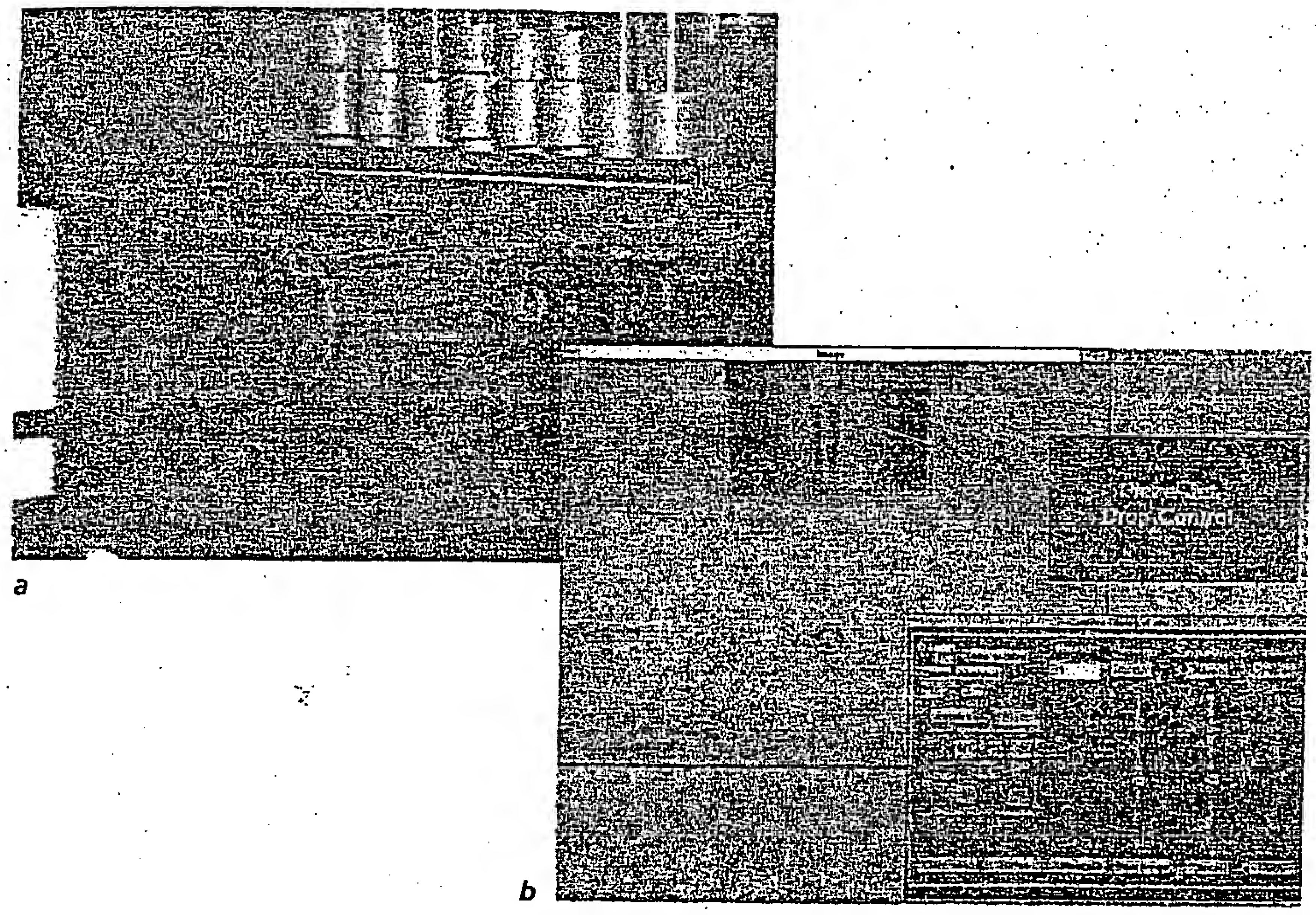


Figure 4 *Automatic drop control. For quality control of drops a stroboscopic image is acquired. (a) Shows the stroboscopic light and the microscope objective connected to the CCD camera. The eight dispensing capillaries can be inspected and adjusted. The drop parameters can be adjusted online (b) using user-friendly buttons for voltage, pulselength and stroboscope delay*

camera is attached to the dispensing head. It enables via software to identify certain cavities on a silicon surface and to dispense a chosen number of droplets independently into each of the chosen cavities (Fig. 5).

The spot size of a microdrop system on a nylon membrane varies between 50 µm and 120 µm and the array density is approximately 2000 spots/cm^2. The functionality of the system allows to dispense on the fly and it takes less than 3 min to array 100 × 100 spots in a square with dot size of 100 µm diameter and 230 µm distance between centres (Fig. 6). At this density it is possible to immobilise a small cDNA library consisting of 14000 clones on a microscope-slide surface.

Another application of the microdispensing technology is the preparation of probe plates for MALDI (Matrix Assisted Laser Desorption Ionisation)-Mass-spectrometry, which has the potential for high throughput applications in DNA analysis. A commercially available MS instrument has a potential to

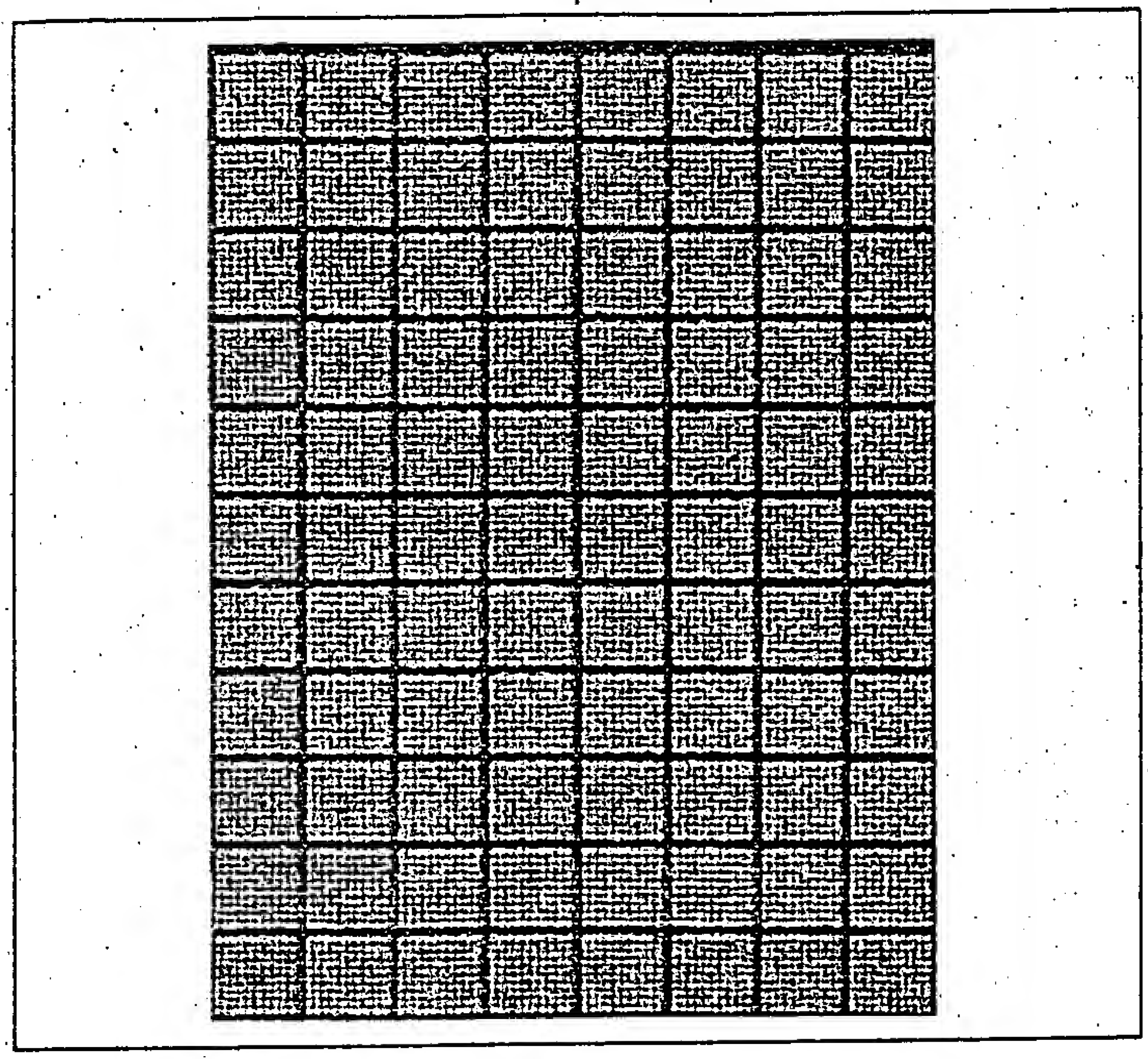


Figure 5 *Piezo Drop Arraying. The picture shows a density of 2150 clones per square centimeter arrayed by the microdispensing device. The spots are 100 μm in diameter and the spacing between the spots is 230 μm. Every spot contains the same DNA fragment. The picture has been taken with a laser scanning device developed at the Max Planck Institute for Molecular Genetics in Berlin*

analyse one probe in a few milliseconds. We have developed a prototype that uses microdispensed arrays of several thousand DNA fragments or proteins on a MALDI-MS target with a size of approx. 2 × 2 cm.

Unfortunately, there are some drawbacks limiting the use of MALDI-MS in genome analysis. For proteins and genome projects the probes have to be purified. Salts and detergents, e.g. SDS out of SDS-PAGE gels or staining reagents, drastically increase the background during the measuring. The background noise overlays the overall spectrum, making it very often impossible to interpret the obtained peaks. Other limitations include the mass range of the investigated species. At the current stage in MALDI-MS DNA sequencing a maximum of 80 bp can be resolved at a one basepair level (Murray, 1996; Kirpekar, Nordhoff et al., 1995).

With the ongoing miniaturisation process in genome analysis, new tools have to be developed for all necessary handling steps in, e.g. a miniaturised

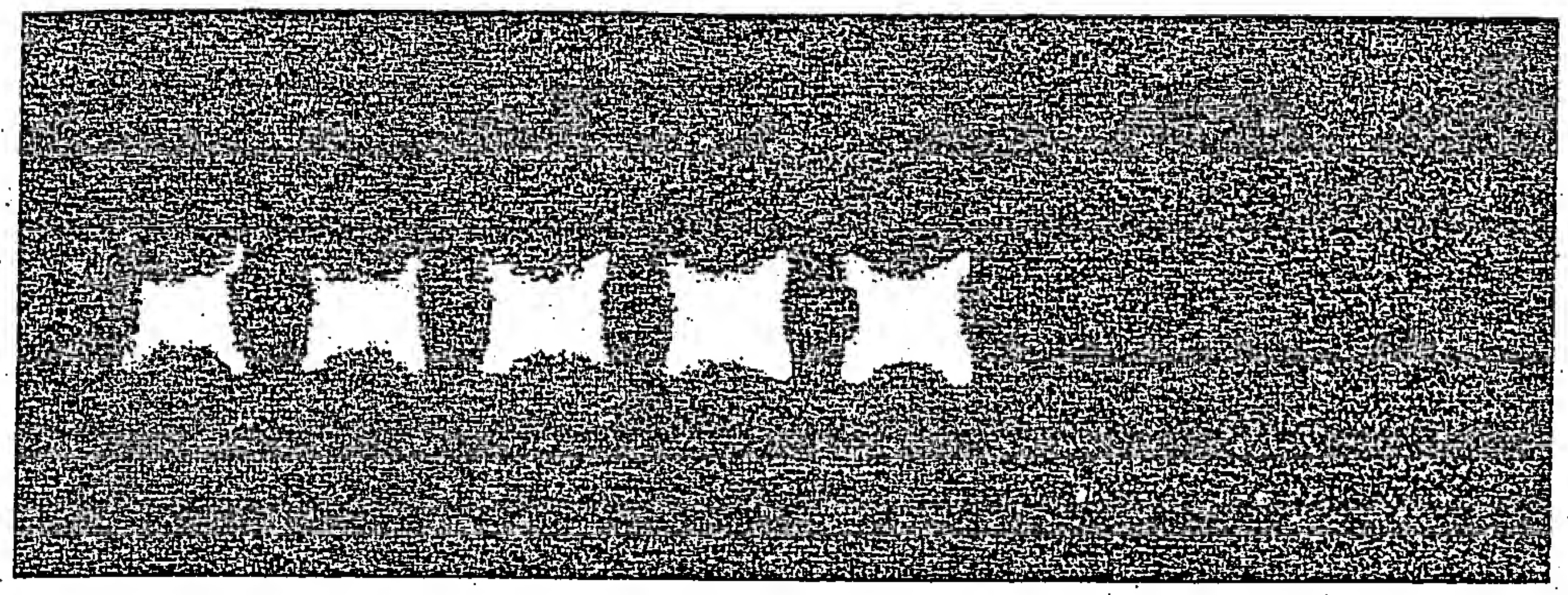


Figure 6 Filling of silicon wafer cavities. Each cavity was anisotropically etched to a size of 0.5 mm × 0.5 mm. The picture shows a row. Five cavities on the lefthand side were filled with 5 nl (50 drops each containing 100 pl) of a fluorescent dye solution (Cy5, Amersham), one cavity was not filled and two cavities on the righthand side were filled with a 10-fold lower concentrated solution of the dye. The total amount of Cy5 in the cavities varies between 1 femtomol (bright signals) and 100 attomoles (weak signals). Picture has been taken with the same detection device as in Figure 5

hybridisation approach. In addition, the detection systems have to be improved. Smaller spot sizes result in less amounts of targets and sensitive detection systems are required with an increased spatial resolution, e.g. using light-optical principles. Optical methods like laser scanning devices for large areas or different microscopy methods including confocal laser scanning microscopes for areas smaller than a few mm^2 are the methods of choice in the near future. For the analysis of single molecules in small cavities other optical methods like scanning nearfield optical microscopy (SNOM) (Moers et al., 1996; Iwabuchi et al., 1997) or fluorescence correlation spectroscopy (FCS) (Rigler, 1995; Oehlenschlager et al., 1996) as well as non optical methods like atomic force microscopy (AFM) (Hansma et al., 1996; Lyubchenko and Shlyakhtenko, 1997) have successfully been tested for individual experiments.

2.4 Conclusions

In the future there will be more and more "Lab-On-A-Chip" devices (Service, 1995), powerful integrations of microfluidics, micromechanics and detection systems (Kovacs et al., 1996). Some application of these devices could be in

the field of faster diagnostics. The microsystems might also include oligonucleotide arrays on different surfaces for DNA diagnostics (Mirzabekov, 1994). There are promising examples of PCR integration on silicon wafers with online detection and/or analysis (Woolley et al., 1996; Taylor et al., 1997). These technologies can screen many DNA samples cheaper, faster and highly parallel. In addition, reducing the size of the processes can provide a new experimental design (Burke et al., 1997), e.g. different techniques of liquid handling like electroosmotic pumps (Freaney et al., 1997) for handling of tiny reaction volumes.

For these scenarios to become reality soon, molecular biologists must work more closely with engineers, physicists and chemists to remove the gap between the macroscopic "laboratory world" and the microsystems "chip world".

References

Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD (1994) Molecule Biology of the Cell. Garland Publishing, New York, London

Burke DT, Burns MA, Mastrangelo C (1997) Microfabrication technologies for integrated nucleic acid analysis [Review]. *Genome Research* 7(3): 189–197

Cantor CR (1990) Orchestrating the human genome project. *Science* 248(4951): 49–51

Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, Winkler J, Lockhart DJ, Morris MS, Fodor SP (1996) Accessing genetic information with high-density DNA arrays. *Science* 274(5287): 610–614

Cherry JL, Young H, Di Sera LJ, Ferguson FM, Kimball AW, Dunn DM, Gesteland RF, Weiss RB (1994) Enzyme-linked fluorescent detection for automated multiplex DNA sequencing. *Genomics* 20(1): 68–74

Dulbecco R (1986) A turning point in cancer research: sequencing the human genome. *Science* 231(4742): 1055–1056

Freaney R, McShane A et al (1997) Novel instrumentation for real-time monitoring using miniaturized flow systems with integrated biosensors. *Annals Clin Biochem* 34(Part 3): 291–302

Geman S, Geman D (1984) Stochastic Relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions in Pattern Analysis and Machine Intelligence*

Hacia JG, Brody LC, Chee MS, Fodor SP, Collins FS (1996) Detection of heterozygous mutations in BRCA1 using high-density oligonucleotide arrays and two-colour fluorescence analysis [see comments]. *Nature Genetics* 14(4): 441–447

Hansma HG, Revenko I, Kim K, Laney DE (1996) Atomic force microscopy of long and short double-stranded, single-stranded and triple-stranded nucleic acids. *Nucleic Acids Res* 24(4): 713–720

Iwabuchi S, Muramatsu H, Chiba N, Kinjo Y, Murakami Y, Sakaguchi T, Yokoyama K,

also include oligo-
ɔstics (Mirzabekov,
n on silicon wafers
Taylor et al., 1997).
er, faster and highly
can provide a new
echniques of liquid
for handling of tiny

ɩiologists must work
:emove the gap be-
nicrosystems "chip

A et al (1997) Novel
ɔr real-time monitoring
flow systems with inte-
Annals Clin Biochem
2
1984) Stochastic Relaxa-
utions and the Bayesian
;es. *IEEE Transactions in*
ıd Machine Intelligence
:, Chee MS, Fodor SP,
)etection of heterozygous
GA1 using high density
ays and two-colour fluo-
[see comments]. *Nature*
-447
o I, Kim K, Laney DE
:e microscopy of long and
ded, single-stranded and
:leic acids. *Nucleic Acids*

ıu H, Chiba N, Kinjo Y,
aguchi T, Yokoyama K,

Tamiya E (1997) Simultaneous detection of near-field topographic and fluorescence images of human chromosomes via scanning near-field optical/atomic-force microscopy (SNOAM). *Nucleic Acids Res* 25(8): 1662–1663

Kirpekar F, Nordhoff E, Kristiansen K, Roepstorff P, Hahner S, Hillenkamp F (1995) 7-Deaza purine bases offer a higher ion stability in the analysis of DNA by matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun Mass Spectrometry* 9(6): 525–531

Kovacs GTA, Petersen K, Albin M (1996) Silicon micromachining – sensors to systems. *Analyt Chem* 68(13): A407–A412

Lehrach H, Bancroft D, Maier E (1997) Robotics, computing, and biology: an interdisciplinary approach to the analysis of complex genomes. *Interdisciplinary Science Reviews* 22: 37–44

Lyubchenko YL, Shlyakhtenko LS (1997) Visualization of supercoiled DNA with atomic force microscopy *in situ*. *Proc Natl Acad Sci USA* 94(2): 496–501

Maier E (1995) Robotic technology in library screening. *Lab Robotics Automation* 7(3): 123–132

Maier E, Crollius HR, Lehrach H (1994) Hybridisation techniques on gridded high density DNA and *in situ* colony filters based on fluorescence detection. *Nucleic Acids Res* 22(16): 3423–3424

Maier E, Meier-Ewert S, Ahmadi AR, Curtis J, Lehrach H (1994) Application of robotic technology to automated sequence fingerprint analysis by oligonucleotide hybridisation. *J Biotech* 35(2-3): 191–203

Maier E, Meier-Ewert S, Bancroft D, Lehrach H (1997) Automated array technologies for gene expression profiling. *Drug Disc Today* 2(8): 315–324

Meier-Ewert S, Maier E, Ahmadi A, Curtis J, Lehrach H (1993) An automated approach to generating expressed sequence catalogues. *Nature* 361(6410): 375–376

Meinkoth J, Wahl G (1984) Hybridization of nucleic acids immobilized on solid supports. *Analyt Biochem* 138(2): 267–284

Mirzabekov AD (1994) DNA sequencing by hybridization – a megasequencing method and a diagnostic tool. *Trend Biotechnol* 12: 27–32

Moers MH, Kalle WH, Ruiter AG, Wiegant JC, Raap AK, Greve J, de Grooth BG, van Hulst NF (1996) Fluorescence *in situ* hybridization on human metaphase chromosomes detected by near-field scanning optical microscopy. *J Microscopy* 182(Pt 1): 40–45

Murray KK (1996) DNA sequencing by mass spectrometry. *J Mass Spectrometry* 31(11): 1203–1215

Oehlenschlager F, Schwille P, Eigen M (1996) Detection of HIV-1 RNA by nucleic acid sequence-based amplification combined with fluorescence correlation spectroscopy. *Proc Natl Acad Sci USA* 93(23): 12811–12816

Rigler R (1995) Fluorescence correlations, single molecule detection and large number screening. Applications in biotechnology. *J Biotech* 41(2–3): 177–186

Saiki RK, Bugawan TL, Horn GT, Mullis KB, Erlich HA (1986) Analysis of enzymatically amplified beta-globin and HLA-DQ alpha DNA with allele-specific oligonucleotide probes. *Nature* 324(6093): 163–166

Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N (1985) Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* 230(4732): 1350–1354

Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW (1996) Parallel Human Genome analysis – microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci USA* 93(20): 10614–10619

Service RF (1995) The incredible shrinking laboratory. *Science* 268(5207): 26–27

Taylor TB, Winn-Deen ES, Picozza E, Woudenberg TM, Albin M (1997) Optimization of the performance of the polymerase chain reaction in silicon-based microstructures. *Nucleic Acids Res* 25(15): 3164–3168

Tilghman SM (1996) Lessons learned, promises kept – a biologist's eye view of the Genome Project. *Genome Research* 6(9): 773–780

Volkmuth WD, Duke T, Austin RH, Cox EC (1995) Trapping of branched DNA in microfabricated structures. *Proc Natl Acad Sci USA* 92(15): 6887–6891

Watson JD, Crick FHC (1953) Molecular structure of nucleic acids. A structure for deoxyribose nucleic acid. *Nature* 171: 737–738

Wetmur JG (1991) DNA probes: applications of the principles of nucleic acid hybridization. *Crit Reviews Biochem Molecular Biol* 26 (3-4): 227–259

Woolley AT, Hadley D, Landre P, Demello AJ, Mathies RA, Northrup MA (1996) Functional integration of PCR amplification and capillary electrophoresis in a microfabricated DNA analysis device. *Analyt Chem* 68(23): 4081–4086

Zehetner G, Lehrach H (1994) The reference library system – sharing biological material and experimental data. *Nature* 367(6462): 489–491

JOURNAL OF BIOMOLECULAR SCREENING
Volume 3, Number 4, 1998

Comparison of Automation Equipment in High Throughput Screening

MICHELLE E. STEVENS, PETER J. BOUCHARD, ILONA KARIV, THOMAS D.Y. CHUNG, and KEVIN R. OLDENBURG

ABSTRACT

This report compares several types of liquid handling equipment presently used in HTS. The devices include 96-well automated pipettors such as the Carl Creative PlateTrac™ (Harbor City, CA), Matrix PlateMate™ (Hudson, NH), Tomtec Quadra-96™ (Hamden, CT) and a Zymark RapidPlate-96™ (Hopkinton, MA) integrated into a full robotic system. A general set of considerations including ease of programming, assay-completion time, accuracy and precision of liquid dispensing, and low-volume pipetting were evaluated. Both a protease screen and a cell-based reporter gene assay were used as model systems for comparison. The data indicate that the Carl Creative PlateTrac has an advantage in several areas. These include the ease in programming, reduction in assay run time, and increased accuracy and precision in liquid dispensing, especially for volumes of 1 μl or less. However, both the Matrix Platemate and Zymark robotic systems may be used to perform complicated multi-step tasks involving multidirectional plate transfer, which is not possible on the current PlateTrac. Advantages and limitations of each piece of equipment are discussed further in this report.

INTRODUCTION

HIGH THROUGHPUT SCREENING (HTS) is an essential component in the discovery of therapeutically active small molecules. Some key considerations in efficient HTS automation are (1) the speed at which the system can process plates, (2) the accuracy and precision of the liquid handling, (3) the ease of equipment programming and operation, and (4) the ability to pipet small volumes (<1 μl). Although the majority of the equipment used today in HTS allows simultaneous 96-well pipetting, the increase in compound library size provides a challenge to develop and use the most efficient automation for screening.

The main criteria for analyzing equipment efficiency in this study were assay programming time, assay run time, minimum dispense volume, and accuracy and precision of pipetting of both organic and aqueous solutions. Dimethylsulfoxide (DMSO) and water were selected as typical examples of each solution type. The rationale for including programming time in the equipment comparison criteria is that for biological assays that require multiple steps, instrument programming can be time consuming and require highly trained personnel. Plate manipulation and reagent addition time are the key limiting factors in the number of plates that can be processed in a given assay run. Higher throughputs can be achieved with more efficient and expeditious liquid handling workstations or systems. Accurate and precise liquid dispensing significantly decreases intra- and interplate variation. In addition, the ability to accurately transfer small volumes of test compound decreases the number of pipetting steps because compound dilution steps are minimized or eliminated entirely. For example, in cell-based assays where sensitivity to organic solvents such as DMSO is a limiting factor, the capability to transfer 1 μl or less of compound in neat DMSO to the assay is essential.

For further equipment validation and comparison, two commonly used HTS assays, a protease assay and whole cell-based reporter gene assay, were chosen. The protease assay selected is well suited for evaluation of the criteria noted above for several reasons. The assay is simple and easily adaptable to HTS and the automation evaluated in this article. The enzyme and related reagents remain stable for the length of the study. Finally, it uses fluorescence as the detection method, which is a well-established signal type that remains linear with concentration. In contrast, cell-based assays, especially transcriptional assays using reporter gene based systems, provide a unique challenge in adaptation to the HTS format.[1,2] Accurate and ef-

DuPont Pharmaceuticals, Experimental Station, Wilmington, DE 19880-0400.

ficient automation can be the key to whether a cell-based assay is scalable for true HTS. The cell-based assay selected utilizes a luciferase reporter system, which is highly sensitive and generates a signal that is stable for 5 hours after substrate (LucLite™, Packard Instrument Company, Meridan, CT) addition.

MATERIALS AND METHODS

Accuracy and precision

The pipetting accuracy was evaluated by pipetting a given volume in the range of 0.5 to 100 μl into each well of a 96-well plate and measuring the change in mass of the entire plate. Ultimately, the change in plate mass in conjunction with the known pipetting precision yields a measure of pipetting accuracy. The precision of liquid handling for each instrument was determined by measuring the coefficient of variability (%CV) across a five-plate set. More specifically, the well-to-well precision was determined by pipetting 7-methoxycoumarin-4-yl acetic acid (MCA) (BACHEM, Torrance, CA), either in water or neat DMSO into Packard black HTRF plates (Packard Instrument Company, Meridan, CT). Plates were then read on a Cytofluor series 4000 fluorometer (Perseptive Biosystems, Framingham, MA). Source plates were made with stock concentrations of MCA of 1 to 200 μM in 200 μl of distilled water or neat DMSO. The test plates were then prepared by transferring the desired volume from the source plate directly into each of five separate test plates. The same set of pipette tips was used for each five-plate set at each volume. The final volume in the test plates was brought to 100 μl with distilled water or DMSO (consistent with the solution type transferred) to yield a final concentration of 1 μM MCA and the plates were then read at Ex_{313nm}, Em_{395nm}.

The pipetting protocols used for each instrument were similar. The manufacturer-designed tips were used for each instrument. For liquid handling on the Tomtec Quadra-96™ (Hamden, CT), Matrix PlateMate, and the Zymark Rapid-Plate-96 an air gap was first pulled, the actual liquid volume was aspirated, and the entire volume aspirated was then dispensed. On the Carl Creative PlateTrac, a 1-μl liquid pre-dispense and 1-μl liquid carry volume were used. The pipette tip touch to the plate was variable, depending on the volume transferred. At 0.5 and 1 μl, the tips were programmed to touch the well bottom for all instruments. For the larger volumes, the liquid transfer was performed at user-programmed heights above the well bottom. A tip touch was not used after dispense.

Matrix metalloprotease/MCA assay

The intra- and interinstrument reproducibility for all four instruments were evaluated using a matrix metalloprotease (MMP) assay as follows. Plates were set up with eight control wells, an eight-point titration curve of a known inhibitor (in duplicate), and 72 wells of DMSO as an indicator of variation across the plate. Assay plates were first prepared by pipetting 1 μl of sample at 100 times the desired final concentration into dry Packard black HTRF plates. A quantity of 84 μl of MCA peptide substrate (10 μM final concentration) in 1× reaction buffer (50 mM TRICINE, 10 mM $CaCl_2$, 0.002% NaN_3, pH 7.5) was then added to each well. The MCA peptide has the sequence Mca-Pro-Leu-Gly-Leu-Dpa-Ala-Arg-NH_2 (Mca = 7-methoxycoumarin acetic acid; Dpa = 3(2′4′-dinitrophenyl)-2,3-diamino propionic acid). Finally, 15 μl of MMP enzyme (12.5 nM final concentration) was added to each well for a total assay volume of 100 μl. The enzyme was thawed just prior to use and diluted to 83 nM with enzyme buffer (50 mM TRICINE, 0.05% BRIJ-35, 400 mM NaCl, 10 mM $CaCl_2$, 0.02% NaN_3, pH 7.5). Plates were incubated for 90 min at 25°C on an orbital shaker. After incubation, 20 μl/well EDTA (500 nM) was added to quench the reaction and the plates were read on a Cytofluor 4000 series fluorometer at Ex_{313nm}, Em_{395nm}.

Cell-based reported gene assay

The PlateTrac and the Zymark integrated robotic system were selected for comparison of reproducibility using a cell-based luciferase reporter gene assay. The PlateTrac can handle all aspects of this assay except pipetting of cells. The cells used in the assay would not remain in uniform suspension throughout the run, so the cells were instead added using a Titertek multidrop (Titertek Instruments, Inc., Huntsville, AL). Although the Zymark robotic system can handle all aspects of the cell-based assay, in the interest of efficiency the choice was made to use the PlateTrac for compound preparation. In addition, the assay plates were incubated in the absence of CO_2 in the Zymark run because frequent incubator door access yielded inconsistent CO_2 exposure throughout the assay run.

The PlateTrac device was used to prepare all assay plates with the test compounds. For a cell-based assay it is critical to maintain consistent DMSO levels due to the sensitivity of mammalian cells to organic solvents. From each compound source plate, 1.25 μl of compound per well was pipetted directly into the appropriate assay plate at 100× final concentration. For screening using the PlateTrac, black Packard HTRF plates were used to reduce the nonspecific background fluorescence detected on the Packard TopCount™. For on-line screening using the Zymark integrated robotic system, white Packard Optiplates were used in conjunction with the Packard LumiCount, already integrated into the Zymark system, in order to maximize signal.

Two stably transfected Jurkat T-cell lines were used in this assay: an inducible expressor of the gene of interest, used to determine specific activity of the tested compounds, and a constitutive expressor used to determine cell suppression/toxicity. These cells were routinely maintained in RPMI 1640 (Gibco BRL, Gaithersburg, MD) with 5% FBS (Hyclone), 1× nonessential amino acids (Gibco BRL), L-glutamine, phenol red, and 3 mg/ml Geneticin (Gibco BRL) as a selection agent to ensure continued expression of their construct. Cell preparation techniques were identical for both clones. On the day of the assay, the cells were prepared in RPMI 1640 without phenol red (Gibco BRL), 5% FBS, 1× nonessential amino acids, 200 μM HEPES, and 0.1% Gentamycin (Gibco BRL), and resuspended to a concentration of 8×10^5 cells/ml to give 50,000 cells/well final concentration in the assay. Media without phenol red must be used in the assay system due to interference in readout of the luminescence by the phenol red. Addition of HEPES ensured consistent pH throughout the assay run.

For the assay on the PlateTrac, 61 μl of the media with or without activators was added to each well of a compound plate. Using the Titertek Multidrop dispenser, 65 μl of cell solution was pipetted into each well of all plates. The plates were incubated at 37°C in a humidified incubator (+5% CO_2) for 5 hr. Using the PlateTrac system for reagent addition, 125 μl of LucLite™ reagent was then added to each well of all the plates. The use of LucLite allowed measurement of luminescence as a glow reaction with a stable signal duration of 5 hr. The plates were read for luminescence on the Packard Top Count™.

For the Zymark automated system, all plates and reagents were prepared and placed in the proper locations prior to commencing the run. The Zymark robotic system then retrieved the assay plate (containing 1.25 μl compound) from the incubated carousel. A total of 61 μl media with or without activators was then added to the appropriate plate wells using the Reagent Addition Station (RAS) and the Zypettor. Cells were added in 65 μl to each well using the Zypettor pipetting station. The plate was then placed in a 37°C humidified cell culture incubator (CO_2) for 5 hr. After incubation, 125 μl LucLite was added to the entire plate using the RAS. The plate was then incubated at room temperature with gentle shaking for 5 min. Plates were read for luminescence on the Packard LumiCount.

RESULTS AND DISCUSSION

Assay adaptation to HTS

In this study the Carl Creative PlateTrac, Matrix PlateMate, Tomtec Quadra-96, and Zymark integrated robotic system were compared. Design of the above equipment, except for the PlateTrac and its use in HTS, have been described elsewhere.[3–6] The PlateTrac design is based on a high-speed conveyor belt system. Two different PlateTrac systems were used in this study. The first system can handle one stack of 96-well plates, allows for plate washing, turbo drying with compressed air, and reagent addition by two separate dispense heads. The second PlateTrac system can handle two stacks of plates, either of which can be the standard 96-well plate or deep well 96-well plates. The system also contains a plate washer and a dispense head with tip wash.

Design differences of each system led to distinct programming requirements for each instrument. The PlateTrac with two dispense heads can run the protease assay, with both reagent additions, as one program. The Tomtec Quadra-96 and the Matrix PlateMate have only one dispense head, which required two programs to be written and run consecutively (or linked) to complete the assay. The Zymark RapidPlate-96 has only one pipetting head, but because it is only one component of an integrated track-based system with a robotic arm, many reagent stations can be used to run an entire assay protocol as one program.

In many cases a successful high throughput screen depends on the ability to transfer small volumes of liquid accurately and precisely from one plate to another. Therefore, the initial set of experiments compared the liquid transfer capability of each device in the range of 0.5 to 100 μl. All experiments were performed in 96-well microtiter plates. For comparison, the capability of each device to pipet both DMSO, which was used to mimick compound transfer, and distilled water, which was used to mimick assay reagent transfer, was tested (Tables 1 and 2). The data indicate that in the low-volume range the PlateTrac repeatedly gave the most accurate and precise values. All of the instruments consistently pipetted DMSO better than water. This is likely due to the differences in surface tension of the two liquids and their interaction with the plastic tips or the recipient plate. Overall, the data indicate that for DMSO solu-

TABLE 1. PIPETTING ACCURACY

	μl Transferred (n = 5 plates)					
	0.5	*1.0*	*2.0*	*5.0*	*10.0*	*100.0*
CarlCreative PlateTrac						
DMSO	0.57 ± 0.11	0.99 ± 0.01	2.03 ± 0.01	5.01 ± 0.03	9.99 ± 0.02	99.93 ± 0.07
Water	0.62 ± 0.17	1.09 ± 0.09	2.06 ± 0.06	5.06 ± 0.04	10.12 ± 0.08	100.08 ± 0.12
Matrix PlateMate						
DMSO	0.61 ± 0.23	1.09 ± 0.01	2.05 ± 0.01	4.82 ± 0.04	10.1 ± 0.02	100.09 ± 0.01
Water	ND[a]	0.62 ± 0.14	1.58 ± 0.64	4.38 ± 0.25	9.85 ± 0.32	99.87 ± 0.11
Tomtec Quadra-96						
DMSO	ND[a]	1.59 ± 0.07	2.33 ± 0.05	4.98 ± 0.05	9.84 ± 0.07	94.40 ± 0.37
[b]						**99.29 ± 0.16**
Water	ND[a]	0.62 ± 0.05	1.42 ± 0.07	3.90 ± 0.07	8.73 ± 0.04	93.13 ± 0.03
[b]		**0.86 ± 0.14**	**1.71 ± 0.04**	**5.04 ± 0.07**	**9.99 ± 0.13**	**99.89 ± 0.13**
Zymark RapidPlate-96						
DMSO	ND[a]	1.09 ± 0.09	2.05 ± 0.10	4.93 ± 0.13	9.66 ± 0.12	97.10 ± 0.16
[b]						**100.32 ± 0.34**
Water	ND[a]	0.58 ± 0.10	1.38 ± 0.03	4.23 ± 0.12	8.82 ± 0.02	95.97 ± 0.37
[b]	**0.58 ± 0.10**	**1.38 ± 0.03**	**2.43 ± 0.11**	**5.18 ± 0.07**	**10.01 ± 0.06**	**100.03 ± 0.26**

Pipetting accuracy determined by dispensing set volumes of either DMSO/MCA solution or water/MCA solution into preweighed plates. The net weight of fluid dispensed into each plate was recorded, divided by the density of either the water or DMSO at room temperature, and divided by 96, multiplied by 1000 to obtain average microliters per well.

[a]Programming limitations did not allow for programming volumes in less than 1 μl.

[b]Actual volume transferred after nominal volume settings were adjusted to yield values closest to desired volume.

TABLE 2. AUTOMATION PIPETTING PRECISION (%CV)

	Coefficient of variation (n = 480 wells)					
	0.5	*1.0*	*2.0*	*5.0*	*10.0*	*100.0*
CarlCreative PlateTrac						
DMSO	4.05	2.4	2.4	1.8	1.5	1.26
Water	8.3	6.7	4.5	3.0	2.0	1.67
Matrix PlateMate						
DMSO	11.2	7.4	3.9	4.5	4.1	2.0
Water	ND[a]	11.5	8.6	7.6	6.5	3.5
Tomtec Quadra-96						
DMSO	ND[a]	3.1	2.9	2.2	2.0	2.3
Water	ND[a]	12.5	4.8	2.0	2.1	2.9
Zymark RapidPlate-96						
DMSO	ND[a]	5.1	5.4	2.9	1.8	2.2
Water	10.9	4.0	2.9	2.3	3.4	1.1

Pipetting precision was determined by dispensing solutions of either DMSO/MCA or water/MCA into dry Packard black HTRF plates. The wells were then brought to a final volume of 100 μl with water using the same dispense head and plates were read on a Cytoflour series 4000 flourometer at Ex_{313nm}, Em_{395nm}. The means, standard deviations, and % CVs (standard deviation divided by the mean and expressed as a percentage) were calculated for all rows and columns of the 96-well plate.

[a]Programming limitations did not allow for programming volumes in less than 1 μl.

tions, the PlateTrac also had the lowest CV across the entire liquid dispensing range. For aqueous solutions, the Zymark RapidPlate-96 (in most cases) yielded the lowest CVs. It is important to note that in the case of the Quadra-96 and RapidPlate-96, two values are shown for a given nominal pipetting volume with respect to accuracy of actual transferred volume (Table 1). All four instruments were first programmed with the nominal volume desired to be pipetted and the results expressed accordingly. Volume settings were then programmed to yield more accurate results (i.e., closer to the desired volume) when necessary. The resultant actual volumes transferred are shown in bold for Table 1. The disparity between nominal and real volumes is an important consideration, particularly when an instrument may be out of calibration or is an aging member of an automated pipetting inventory. The Matrix PlateMate used for the low volumes (10 μl and less) was a demonstration instrument brought in for this comparison, and the Tomtec Quadra-96 is the oldest of the four systems evaluated. These results stress the importance of determining the accuracy of an automated pipetting system when validating any assay. Because data quality is proportional to pipetting accuracy and precision, reagent properties vary greatly and can significantly impact results as shown in Table 1. At the higher volume of 100 μl, all of the instruments were very precise, with CVs approximately 3% or less. For all volumes tested, all instruments pipetted within the manufacturers reported specifications.

The time required to program each instrument and the potential assay throughput were compared (Table 3). The data indicate that the PlateTrac was the fastest to program and also had the quickest plate-handling cycle time. The fast programming time was due to the touch screen program control, which allowed rapid program creation in conjunction with the validation for plate heights, speed controls, and working volume range. The disadvantage of this type of programming was that only relatively simple programs could be entered. For example, sequential applications, such as dilution and transfer, must be written as two separate programs. The Tomtec Quadra-96 and the Matrix PlateMate were comparable in programming and plate cycle time. The Zymark System required longer initial programming time, but once set up, it was fully automated. The other three instruments are workstations that must be attended to when stacker capacities have been reached. The PlateTrac allowed the fastest plate handling of 0.5 min for plate destack, reagent aspiration and dispense, and plate restacking, as compared with over 2 min for the same process for the other instruments. The very fast plate cycle time for liquid handling was made possible by the conveyor belt design, allowing separate modules of the machine to be working at the same time on different plates. The very fast cycle time would make the PlateTrac the preferred choice where rapid addition of a stable reagent is desired to extend the total throughput of a process.

TABLE 3. MMP-3 APPROXIMATE PROGRAMMING AND LIQUID HANDLING CYCLE TIMES

	Programming time	*L/H cycle time/plate*
CCWS PlateTrac	10 min	0.5 min
Matrix PlateMate	35 min	2.3 min
Tomtec Quadra-96	25 min	2.45 min
Zymark RapidPlate-96	45 min[a]	2.67 min

Programing times were determined for writing the program to perform all required steps on that instrument, and running the program validation for speed settings and plate heights. Times take into account user familiarity with instrument programming. Liquid handling cycle times include time to process plate through program and loading of fresh tips where required.

[a]Denotes entire assay programming time as opposed to solely liquid handling programming time.

TABLE 4. COMPARISON OF IC_{50} VALUES OF A KNOWN MMP-3 INHIBITOR ON EACH INSTRUMENT

	IC_{50} (nM)			
	PlateTrac	*Matrix PlateMate*	*Tomtec Quadra-96*	*Zymark Rapid-Plate-96*
Plate 1	11.85	11.92	15.07	11.58
Plate 2	11.95	12.05	13.76	11.61
Plate 3	11.92	12.03	14.24	12.93
Plate 4	11.93	11.99	14.15	11.78
Plate 5	11.92	11.91	15.79	11.11
Average	11.91	11.98	14.60	11.80

The IC_{50} of a known MMP-3 inhibitor was determined for each plate in a five plate set on each instrument. Each IC_{50} value represents the mean of duplicate well values in the IC_{50} titrations for that plate. Assay was run in 100 μl final volume.

Automation of protease assay in HTS

MMPs are widely involved in extracellular matrix remodeling and therefore comprise an attractive model for various pathologic processes, such as morphogenesis, angiogenesis, and metastasis.[7,8] Due to the wide utility, ease, and number of examples of protease-based assays, an MMP assay was chosen for use as a model system to test the capabilities of the devices mentioned above. In this assay a donor-quencher fluorogenic peptidyl substrate was used. The basis of the assay was the release of the Dpa quench, as a result of the cleavage of the Leu-Dpa bond, generating a fluorescent signal. The specific MMP inhibitor was selected from an in-house compound library.

Five inhibitor titration plate replicates were screened on the PlateTrac, PlateMate, Quadra-96, and Zymark integrated robotic system. The IC_{50} of a known inhibitor (Table 4) and DMSO to mimic random compound distribution (Table 5) were used to determine the intra- and interplate variability. These aspects are critical in the subsequent statistical analysis to identify potentially active compounds. The data indicate that the PlateTrac and the RapidPlate-96 on the Zymark system had the lowest variability both within a single plate and across a five-plate set.

Automation of cell-based assay in HTS

The final set of experiments compared the use of the PlateTrac and the Zymark integrated robotic system in a cell-based assay. The Quadra-96 and the Matrix PlateMate were not included in this set of experiments because the initial data supported the Carl Creative PlateTrac as the most efficient off-line workstation to compare to a fully integrated robotic system. In this study, a luciferase reporter gene fused to four consensus sequence copies of the transcriptional binding site of the gene of interest was used. Two cell lines were used (inducible and constitutive) to allow for the distinction of compounds that displayed true inhibition of the expression of the gene of interest, versus general suppression/toxicity by the compound. The use of luciferase comprised a sensitive and convenient read-out system and was previously validated in the HTS format.[9]

To compare both automation routes, ten random compound plates from an in-house chemical library were screened against both cell lines. All plates also included the following controls: nonactivated cells, activated cells, and wells containing both activated cells and a specific inhibitor. All of the compound activity values were normalized for the activated cell control response and expressed as percentage inhibition. Percentage CVs were similar for both instrument systems and were in the range of 9% to 18% for the PlateTrac and 11% to 21% for the Zymark. However, for the Zymark activity run, the average compound inhibition was 33%, as compared with 0% for the PlateTrac data set (Fig. 1). It is unclear what factors may have contributed to this shift, but activity cutoffs were determined consistently for both data sets. The toxicity assay data distributions were similar for both systems (Fig. 1). The discrepancy between the activity and toxicity data may be due to the difference in the mechanism of luciferase expression for the inducible versus constitutive cell lines. The cell lines also respond to stress at different levels (data not shown), and the lack of

TABLE 5. MMP-3 ASSAY INTRA- AND INTERPLATE VARIABILITY

	Average signal ± SD (%CV)			
	CCWS PlateTrac	*Matrix PlateMate*	*Tomtec Quadra-96*	*Zymark Rapid-Plate-96*
Plate 1	159.2 ± 3.1 (1.9%)	122.9 ± 5.0 (4.1%)	184.9 ± 5.3 (2.8%)	220.6 ± 5.2 (2.4%)
Plate 2	161.3 ± 3.3 (2.0%)	116.8 ± 4.7 (4.1%)	200.0 ± 7.0 (3.5%)	233.4 ± 5.8 (2.5%)
Plate 3	165.6 ± 2.8 (1.7%)	116.5 ± 3.8 (3.8%)	211.5 ± 7.7 (3.6%)	239.6 ± 5.5 (2.3%)
Plate 4	162.5 ± 3.1 (1.9%)	118.9 ± 3.9 (3.3%)	218.3 ± 10.3 (4.7%)	236.3 ± 5.6 (2.4%)
Plate 5	163.8 ± 3.5 (2.1%)	119.8 ± 3.5 (2.9%)	222.9 ± 6.1 (2.7%)	229.4 ± 5.6 (2.5%)
Average	162.5 ± 3.2 (2.0%)	119.0 ± 4.2 (3.6%)	207.5 ± 7.3 (3.5%)	231.9 ± 5.5 (2.4%)

The well-to-well variation was determined within the plate by calculating the mean signal: background, standard deviation, and % CV of predotted DMSO (mimic compound transfer) with substrate and enzyme (n = 72). The interplate variability was determined for each instrument across the five-plate set.

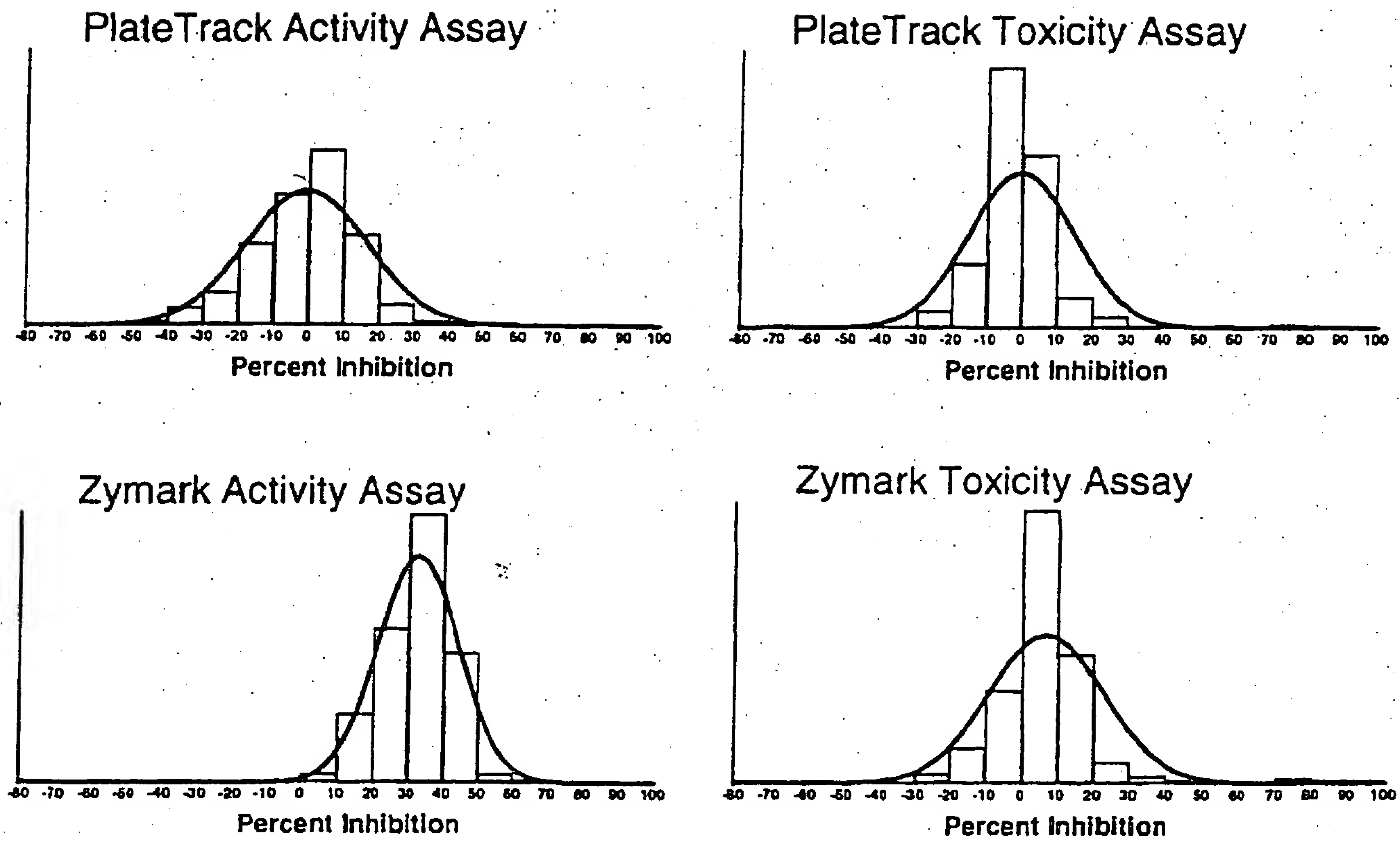


FIG. 1. Distribution across a ten plate random compound set in the cell-based reporter gene assay for the PlateTrac and the Zymark system. The distribution is shown for all ten compound plates screened in each cell line for both the PlateTrac and the Zymark system. The activity data represent the inducible cell line assay and the toxicity data were generated using the constitutively expressing cell line.

CO_2 in the Zymark run may put a higher level of stress on the inducible cell line.

The cutoff values for identifying active compounds in both the activity and toxicity assays were established by calculating the mean percentage activity values plus two standard deviations of the sample wells. Compounds were considered to be active if they were found to yield an activity value greater than two standard deviations above the mean (95.5% confidence). The activity cutoff values are different for each assay on each instrument due to variations in the means and standard deviations determined for each plate set run. For the PlateTrac instrument, the activity cutoff, expressed as percentage inhibition, was 34%, and the toxicity cutoff was 31%. For the Zymark instrument the activity cutoff was 56% and the toxicity cutoff was 40%. The summary of the active compounds is presented in Table 6. From a total of ten compounds identified as active, six were in agreement for both instrument runs to be both active and not toxic. Of the four remaining compounds, two (compounds 6 and 7) were found to be active on the PlateTrac, borderline active on the Zymark, but not toxic for both instruments. Compound 3 was active but not toxic on the PlateTrac only. This same compound was weakly active, but toxic on the Zymark. Finally, compound 9 was inactive on the PlateTrac, active on the Zymark, and not toxic on both instruments. In most cases the percentage inhibition of the tested sample was higher for the Zymark run. The difference may be due to the absence of CO_2 during the assay runs on the Zymark, as discussed above. Confirmatory testing is necessary to further determine activity and toxicity characteristics of these compounds.

TABLE 6. COMPARISON OF COMPOUND HITS IN THE CELL-BASED REPORTER GENE ASSAY ON THE CCWS AND THE ZYMARK SYSTEM

	% Inhibition			
	Activity assay		*Toxicity assay*	
	PlateTrac	*Zymark*	*PlateTrac*	*Zymark*
Compound 1	86	95	0	2
Compound 2	64	58	4	0
Compound 3	57	40	6	60
Compound 4	41	61	0	11
Compound 5	45	82	4	22
Compound 6	43	54	4	14
Compound 7	39	55	0	19
Compound 8	70	100	0	0
Compound 9	17	89	12	0
Compound 10	39	68	0	2
Cutoffs ($\pm 2\sigma$)	34	56	31	40

The compound hits are recorded as percentage inhibition of control wells. The cutoff for a compound hit was determined as the average activity plus 2 SD. The cutoff criteria are different for each instrument due to variability of the plates across a ten-plate set.

CONCLUSION

The PlateTrac is the most accurate and precise instrument of the four tested when pipetting small volumes. This provides an invaluable advantage for compound transfer in any assay. In addition, the PlateTrac system requires less time for programming and has a significantly faster plate processing time when compared with the other instruments. Another advantage of the PlateTrac is that the pipetting heads can be easily interchanged between the 50-μl 96-well format and the 200-μl 96-well format. The Matrix PlateMate, Tomtec Quadra-96, and Zymark integrated robotic system have an advantage over the current PlateTrac in programming multi-step operations requiring multi-directional plate movement (newer versions of the PlateTrac allow multi-directional plate movement but were not available for testing). This is one of the reasons that the programming time is more extensive on the other equipment than for the PlateTrac. The Zymark integrated robotic system is the only instrument tested that has the advantage of full on-line assay automation without human intervention.

Intra- and interplate variability results for the protease assay showed all four instruments to perform well with CVs under 5% and relatively good correlation between instruments. The IC_{50} data for the control compound in this assay was also very reproducible between instruments.

For the cell-based reporter gene assay, both the PlateTrac and the Zymark integrated robotic system provide an efficient design for HTS. However, each system has key limitations. The PlateTrac, with the instrument design tested here, is not efficient for cell suspension and transfer. The Zymark system is limited in the ability to quickly prepare compound test plates for an assay and, in the case of this cell-based assay, unable to provide consistent CO_2 levels during the plate incubation.

In summary, to automate a high throughput screen, there are many considerations to take into account. The first of these is the decision to use an integrated robotic system or to use a workstation approach, automating specific tasks rather than the entire assay. If using the workstation approach, one must then decide which instrument is most suitable for the assay. This report provides information on the advantages and limitations of three different workstations and one integrated robotic system to facilitate that decision.

REFERENCES

1. Wallace, R.W. (1997). High throughput screening. *DDT*. 2:557.
2. Parandoosh, Z. (1997). Cell-based assays. *J. Biomol. Screening*. 2:201.
3. Astle, T.W., Akowitz, A. (1996). Accuracy and tip carryover contamination in 96-well pipetting. *J. Biomol. Screening*. 1:211.
4. Sills, M.A. (1997). Integrated robotics vs. task-oriented automation. *J. Biomol. Screening*. 3:137.
5. Zaayenga, A., Harris, C., Brosius, R., and Wildey, M.J. (1997). Use of the Zymark RapidPlate-96 to perform pipet-in-a-tip assays for high throughput screening. *ISLAR '97 Proceedings*, pp. 189–196.
6. Matrix Platemate. (1996). Automatic dispensing/diluting device, user instruction manual. Matrix Technologies Corporation. Hudson, NH.
7. Wojtowics-Praga, S.M., Dickson, R.B., Hawkins, M.J. (1997). Matrix metalloproteinase inhibitors. *Invest. New Drugs*. 15:671.
8. Douglas, D.A., Shi, Y.E., Sang, Q.A. (1997). Computational sequence analysis of the tissue inhibitor of metalloproteinase family. *J. Protein Chem*. 4:237.
9. Suto, C.M., Ignar, D.M. (1997). Selection of the optimal reporter gene for cell-base high throughput screening assays. *J. Biomol. Screening*. 2:7.

Address reprint requests to:
Kevin R. Oldenburg
DuPont Pharmaceuticals Co.
Experimental Station
Wilmington, DE 19880-0400

E-mail: Kevin.R.Oldenburg@dupontpharma.com